2020-04-14

Role of Cis-regulatory Elements in Transcriptional Regulation: From Evolution to 4D Interactions

Pranitha Vangala
University of Massachusetts Medical School

ROLE OF CIS-REGULATORY ELEMENTS IN TRANSCRIPTIONAL REGULATION: FROM EVOLUTION TO 4D INTERACTIONS

A Dissertation Presented

By

Pranitha Vangala

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 14, 2020

Interdisciplinary Graduate Program

ROLE OF CIS-REGULATORY ELEMENTS IN TRANSCRIPTIONAL REGULATION: FROM EVOLUTION TO 4D INTERACTIONS

A Dissertation Presented
By
Pranitha Vangala

This work was undertaken in the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

Under the mentorship of

_____

Manuel Garber, PhD, Thesis Advisor

_____

Kate A. Fitzgerald, PhD, Member of Committee

_____

A.J. Marian Walhout, PhD, Member of Committee

_____

Zhiping Weng, PhD, Member of Committee

_____

Athma Pai, PhD, Member of Committee

_____

Vijay Sankaran, PhD, External Member of Committee

_____

Elinor Karlsson, PhD, Chair of Committee

_____

Mary Ellen Lane, Ph.D., Dean of the Graduate School of Biomedical Sciences

April 14, 2020

Dedication

This thesis is dedicated to my wonderful family.

A special feeling of gratitude to my loving parents, Dr. Ramakrishna and Mrs. Vijayalakshmi whose words of encouragement ring in my ears forever, my sister Prathyusha for upholding my graduate school dream, my wonderful husband Vineeth who never left my side through this journey and constantly supported me even at personal costs, my most special puppy Anand who spent endless hours with me on the couch while analyzing data, reading papers and writing, and comforting me on my gloomy days with his slimy kisses.

Acknowledgements

This Ph.D. has been a fun and arduous journey with many ups and downs. I have been truly fortunate to work with many wonderful people at the University of Massachusetts medical school. I would like to thank them all in these acknowledgments (in no particular order).

It is a great pleasure to thank my supervisor, Dr. Manuel Garber, for his guidance and support throughout my Ph.D. research, for entrusting me with great responsibility and freedom, and for his constant encouragement and unending challenges. His belief in my capabilities, sometimes going beyond my own expectations definitely brought the best out of me.

Manuel also brought together an excellent group of people in his lab who made every day joyous. Special thanks to Elisa Donnard, Barbara Tabak, Hakan Ozadam, Alper Kurchukal, Alan Derr, Kyle Gellatly, Yuqing Wang, Yuming Cao for brightening my gloomy days, sharing ups and downs of science.

I have been fortunate to get the flavor of the University of Massachusetts Medical School even before joining as a graduate student. I thank Dr. Michael Czech for giving me this wonderful opportunity. He taught me many lessons during my 2-year term and strengthened my goal to pursue Ph.D. Next, I thank the “GERP” team Myriam Aouadi, Michaela Tencerova, Joseph Yawe, and Sarah Nicoloro for all the fun days we had together from doing crazy animal experiments to pouring 100s of percoll gradients.

Next, I would like to thank my fantastic classmates, without whose help, an engineer like me would not be able to come this far in a biomedical sciences program.

My thesis committee, Elinor Karlsson, Kate Fitzgerald, Marian Walhout, Athma Pai and Zhiping Weng have guided me thoughtfully over the past couple of years, and I am grateful for their advice and insight.

Special thanks to Heidi Beberman for her endless support and patience and for making it easy for me to navigate through my Ph.D.

I’m also extremely lucky to have a fantastic scientific community at the University of Massachusetts Medical School.

Abstract

Transcriptional regulation is the principal mechanism in establishing cell-type specific gene activity by exploring an almost infinite space of different combinations of regulatory elements, transcription factors with high precision. Recent efforts have mapped thousands of candidate regulatory elements, of which a great portion is cell-type specific yet it is still unclear as to what fraction of these elements is functional, what genes these elements regulate, or how they are established in a cell-type specific manner. In this dissertation, I will discuss methods and approaches I developed to better understand the role of regulatory elements and transcription factors in gene expression regulation.

First, by comparing the transcriptome and chromatin landscape between mouse and human innate immune cells I showed specific gene expression programs are regulated by highly conserved regulatory elements that contain a set of constrained sequence motifs, which can successfully classify gene-induction in both species. Next, using chromatin interactions I accurately defined functional enhancers and their target genes. This fine mapping dramatically improved the prediction of transcriptional changes. Finally, we built a supervised learning approach to detect the short DNA sequences motifs that regulate the activation of regulatory elements following LPS stimulation. This approach detected several transcription factors to be critical in remodeling the epigenetic landscape both across time and individuals.

Overall this thesis addresses several important aspects of cis-regulatory elements in transcriptional regulation and started to derive principles and models of gene-expression regulation that address the fundamental question: “How do cis-regulatory elements drive cell-type-specific transcription?”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of copyrighted Materials Produced by the Author

Parts of this dissertation have been published, submitted or in preparation for publication as:

• Donnard Elisa*, **Pranitha Vangala***, Shaked Afik*, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, et al. 2018. “Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon.” Cell Systems 6 (3): 381–94.e7. * Co-first authors

• **Vangala Pranitha**, Murphy Rachel, Quinodoz Sofia A., Gellatly, Kyle, McDonel, Patrick, Guttman, Mitchell, Garber Manuel, “High-resolution mapping of multi-way enhancer promoter interactions regulating pathogen detection” (Under Revision)

• Shaked Afik*, **Pranitha Vangala***, Elisa Donnard, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, Patrick McDonel, Jeremy Luban, Manuel Garber, Nir Yosef. The publication will be entitled “Uncovering the short DNA sequences which control the epigenetic landscape of Dendritic cells maturation”. (In preparation) * Co-first authors

Other published works during my graduate study that are not presented in this thesis:

• Kriegsman, Barry A., **Pranitha Vangala**, Benjamin J. Chen, Paul Meraner, Abraham L. Brass, Manuel Garber, and Kenneth L. Rock. 2019. “Frequent Loss of IRF2 in Cancers Leads to Immune Evasion through

Decreased MHC Class I Antigen Presentation and Increased PD-L1 Expression.” Journal of Immunology 203 (7): 1999–2010.

• Nyalwidhe, Julius O., Glen R. Gallagher, Lindsey M. Glenn, Margaret A. Morris, **Pranitha Vangala**, Agata Jurczyk, Rita Bortell, David M. Harlan, Jennifer P. Wang, and Jerry L. Nadler. 2017. “Coxsackievirus-Induced Proteomic Alterations in Primary Human Islets Provide Insights for the Etiology of Diabetes.” Journal of the Endocrine Society 1 (10): 1272–86.

• Nyalwidhe, Julius O., Agata Jurczyk, Basanthi Satish, Sambra Redick, Natasha Qaisar, Melanie I. Trombly, **Pranitha Vangala**, et al. 2020. “Proteomic and Transcriptional Profiles of Human Stem Cell-Derived β Cells Following Enteroviral Challenge.” Microorganisms 8 (2). https://doi.org/10.3390/microorganisms8020295.

• Wang, Yetao, Lawrence Lifshitz, Kyle Gellatly, Carol L. Vinton, Kathleen Busman-Sahay, Sean McCauley, **Pranitha Vangala**, et al. 2020. “HIV-1-Induced Cytokines Deplete Homeostatic Innate Lymphoid Cells and Expand TCF7-Dependent Memory NK Cells.” Nature Immunology 21 (3): 274–86.

• Yuming Cao*, Zhiru Guo*, **Pranitha Vangala**, Elisa Donnard, Ping Liu, Patrick McDonel, Jose Ordovas-Montanes, Robert W. Finberg, Jennifer P. Wang#, and Manuel Garber# (in preparation). “Single cell analysis of upper airway cells reveals host and viral dynamics in influenza-infected adults.”

List of Third Party Copyrighted Material

This thesis contains no third party copyrighted material.

I. CHAPTER I: INTRODUCTION

The human body consists of over 200 different cell types, with each performing a highly specialized function. All cells are derived from a single fertilized oocyte and contain the same blueprint DNA. Lineage specification is achieved by a precise choreography of gene expression, which leads to cell-specific programs. In eukaryotes, gene expression regulation can be achieved at any step starting from transcription to mRNA processing to translation or through post-translational modifications of the proteins. In this dissertation, I focus on understanding how gene expression is regulated at the transcription step.

Less than 2% of the human genome accounts for the coding-sequences (i.e gets processed by cellular machinery into proteins) the rest of the genome has non-coding regions. These non-coding regions are composed of several structural elements (like telomeres, centromeres, etc), cis-regulatory elements, introns, repeats, and pseudogenes. Cis-regulatory elements can be broadly classified into two main groups: promoters (elements proximal to transcription start sites) and enhancers (the distal elements). These regulatory elements are central to gene expression, in that they control cell-type specificity, environmental context, and magnitude of gene expression. These elements act as docking sites for several transcription factors (TFs) and the signal from this combination of TFs will express or repress the expression of a gene. Thus, gene expression regulation at transcription in eukaryotes is mainly due to the interplay between multiple cis-regulatory elements and transcription factors that are bound within

them. Even though cis-regulatory elements are dispersed throughout the genome they are able to communicate with each other via the three-dimensional structure of the chromatin.

I.1. Chromatin Structure

The human genome is 2 meters long, and its packaging into the 10 μm nucleus is achieved through several levels of hierarchical foldings. At the first level, DNA is wrapped around histone octamer that consists of 2 copies of the four histone core proteins (H2A, H2B, H3, and H4) and this structure is called a nucleosome. Nucleosomes are further condensed to form higher-order chromatin structure. This next level of folding ensures that the spatially separated DNA regions can now be in very close proximity to each other in three-dimensional (3D) space. Histones in the nucleosome also undergo a variety of post-translational modifications, usually at the C-terminal or N-terminal ends (called histone tails). The amino acid of the histone tail that is modified and the type of modification it undergoes determines the accessibility of the DNA and chromatin compaction. The packaging of chromatin in eukaryotes serves two major functions of the cells (1) it compacts the genomic DNA so it can fit into the nucleus and (2) the context-specific folding and unfolding of chromatin, modulates the accessibility of the genome to the transcription factors which in turn establish the spatiotemporal gene expression patterns.

Microscopy-based methods have shown that the genome is organized into compartments that have distinct functional properties (van Holde 2012). The

nucleolus, heterochromatin, and euchromatin are the prominent nuclear compartments. The nucleolus is organized around regions that encode ribosomal genes. Heterochromatin is found along the nuclear periphery and consists of inactive or repressed genomic regions. Euchromatin, on the other hand, is organized in the nuclear interior and consists of active genes and regulatory elements. Within euchromatin, genes with similar functions are shown to form sub-compartments. For example, histone genes from various chromosomes preferentially form a cluster as well as the active genes of immunoglobulin heavy chain and the receptors of B and T lymphocytes. Recently this functional organization of the nucleus is thought to be due to a phenomenon called phase-separation. In phase separation, a mixture of macromolecules present at high concentration in a solution separate into distinct phases based on their affinity to form compartments (Palikyras and Papantonis 2019; Sabari et al. 2018).

With the advent of high throughput assays for mapping genome organization, the resolution at which the nuclear organization can be studied increased dramatically. These studies showed that there are sub-compartments in euchromatin called topologically associated domains (TADs) (Nora et al. 2012; Dixon et al. 2012). TADs contain DNA elements that make high-frequency interactions with one another compared to elements in adjacent regions. Specific examples show that TADs not only facilitate regulatory interactions for promoters and enhancers but also insulate the gene promoters from spurious activation by enhancers in the neighboring TADs. But genome-wide depletion of TAD

boundary elements like CTCF, cohesin or genome-wide rearrangement of TAD boundaries showed very modest changes in gene-expression (Ghavi-Helm et al. 2019; Splinter 2006; Schwarzer et al. 2017; Yokoshi, Segawa, and Fukaya 2020).

I.2. Epigenome And Epigenetics

The epigenome is defined as the map of chemical changes observed in the DNA molecule and histone proteins. These changes, in fact, don't modify the underlying DNA sequence but alter the way cellular machinery interprets genomic instructions. Although all cells in our body have the same exact blueprint of DNA, what varies among different cells is the epigenome. Epigenetic modifications are uniquely established during cellular differentiation and can undergo further changes in response to environmental factors. Epigenetic modifications are also inheritable i.e, they can be passed onto offspring during cell divisions in the form of "Epigenetic Memory". An example of trans generational epigenetic memory occurring in humans is "the Dutch hunger winter of 1944". In this study, it was shown that maternal malnutrition is correlated to the susceptibility of metabolic disorders in offspring (Schulz 2010). In another example, epidemiological studies showed a reduced risk of hayfever and asthma in children exposed to endotoxins. Prior exposure to endotoxins also resulted in reduced immune activation (Schuijs et al. 2015) due to sustained nucleosomal remodeling at promoter and enhancer regions. This chromatin remodeling is associated with altered responses, thus providing an epigenetic memory of the first stimulus (Netea 2013; Kim et al. 2019).

The best-studied epigenetic modification is DNA methylation, where methyl groups are attached to DNA molecules by enzymes called DNA methyltransferases. DNA methylation is key to processes like X chromosome inactivation, imprinting, and repression of transposable elements (Phillips and Others 2008). The second most popular epigenetic modification is the modifications to the histone tails in the nucleosomes by histone modifiers. Lysines, serines, and arginines of the histone tail undergo modifications such as methylation, acetylation, ubiquitylation, phosphorylation, and sumoylation. These histone tail modifications affect gene expression by either altering the chromatin compaction or by recruiting different histone-modifying enzymes (Bannister and Kouzarides 2011). Several consortia like Roadmap Epigenome, Encode, International Human Epigenome Consortium, Blueprint Epigenome over the last few years have generated epigenetic maps for many different cell types and states. These efforts resulted in a comprehensive catalog that associates particular modifications with various genomic features. For example, when the lysine 9 residue on histone H3 undergoes trimethylation (H3K9me3) it is correlated with constitutive heterochromatin (Zhou, Goren, and Bernstein 2011). H3K4 trimethylation (H3K4me3) is correlated with promoters, H3K4 monomethylation (H3K4me1) is correlated with enhancers, and acetylation of H3K27 (H3K27ac) is correlated with transcriptional activity. H3K27ac can be found both on promoters and enhancers of actively transcribing genes (Creyghton et al. 2010; Zhou, Goren, and Bernstein 2011). Combinations of

various histone modifications can be used to sub-classify regulatory regions. For example, regions marked with H3K4me1 and H3K27ac are usually classified as “active” enhancers while regions marked with only H3K4me1 are considered “poised” enhancers. In recent years, the catalog of putative enhancers has grown and now the estimate is that there are over one million such elements in the human genome. Throughout this dissertation, I heavily rely on H3K27ac, H3K4me3 and H3K4me1 marked regions to define cis-regulatory elements and their functional state.

I.3. Transcriptional Machinery

While the epigenetic landscape establishes the role of genomic regions in a particular cell state or in response to external factors, the transcriptional machinery, composed of RNA polymerases, transcription factors, and other protein or RNA complexes is required for the actual transcription. Eukaryotic cells have 3 classes of RNA polymerases (RNA-Pol) I, II, and III each playing a major role in transcribing different classes of RNA molecules. All protein-coding genes are transcribed by RNA-Pol-II, while ribosomal RNAs (rRNA) and transfer RNAs (tRNA) are transcribed by RNA-Pol-I and RNA-Pol-III. Even though RNA polymerase is a key component of the transcriptional machinery it is not sufficient to drive transcription. RNA polymerases require general transcriptional factors (GTFs) and other transcriptional factors (TFs) for transcription.

RNA polymerases along with GTFs form a pre-initiation complex (PIC) at promoter regions. Although GTFs and RNA polymerase enzymes comprise the

minimum components of transcription initiation, these factors alone generally lead to low transcriptional activity. However, there is a large class of TFs in the cells that don't belong to the GTF class. These TFs bind to either promoter or enhancers and control the expression of specific genes in specific contexts. In the early 2000s, researchers identified an additional co-factor that was required for particularly RNA-pol-II called the "Mediator complex". Initially, the mediator complex was thought to act as a structural bridge between enhancers and promoters facilitating their interaction (Kagey et al. 2010). But Recent studies showed that the mediator complex has a little role as a structural bridge but influences the formation of PIC by recruiting RNA-Pol-II to promoters and second to act as a bridge between transcription factors and RNA-Pol-II (El Khattabi et al. 2019).

TFs can act as activators of gene-expression or can repress it. Many of the TFs have a DNA binding domain which binds to specific genomic sequences. The genome-wide consensus sequence for every TFs is called transcription factor binding motif. During the "OFF state" nucleosomes are tightly wrapped and the underlying DNA is inaccessible to the TFs. But upon switching to permissive/ open chromatin state, the accessible regions are established. These regions give TFs access to their binding sites. The accessible regions are found around the transcription start sites (TSS) in the promoters and enhancers. In general, accessible regions are about 100 to 1000 base-pairs. TFs are often found to be bound as homo or heterodimers. Also, different combinations of TFs are required

for precise gene-expression. These combinations vary between cell-types and cellular context, giving cells another layer to fine-tune gene expression. An example of combinatorial regulation at the TF level can be seen at the Interferon-beta enhancer. This enhancer requires binding of three different TFs NFKB, IRF3 (homodimer) or IRF3/IRF7 (heterodimer) and ATF2/cJUN (heterodimer) upon viral infection to transcribe the IFN-beta gene upon viral infection (Thanos and Maniatis 1995). Often such cases where the assembly of many TFs is required are referred to as “Enhanceosomes”. In enhanceosomes, TFs act in cooperation and the loss of even one bound TF disrupts the function of the enhancer. “Billboard” enhancers, on the other hand, are modular, where the binding of each TF is not necessary for enhancer activity but rather has an additive or synergistic effect (Arnosti and Kulkarni 2005).

I.4. Assays To Map Transcriptional and Epigenetic States

With the advent of sequencing technologies, global maps of RNA and epigenome are available for various different cell types. In recent years the sequencing platforms have been revolutionized and become very affordable. For example, the human genome project which was started in 1990, cost over a billion dollars and took over a decade to finish, while today we can sequence a whole genome for about a thousand dollars in one day (Sheikh 2018). In this section, I will go through the existing assays that are widely used to map different epigenetic and transcriptional states of the cells.

I.4.I. Transcriptome Profiling by RNA-Seq

RNA-Seq Is a Transcriptome Profiling Technique That Uses Next-Generation Sequencing. RNA-Seq Was First Done in 2007 To Capture Poly(a) mRNAs. The Method Commonly Involves Isolation of Total RNA From the Cell(S) and Reverse Transcription To Generate cDNA. Finally, DNA Adapters Are Ligated To Enable Sequencing on Next-Generation Sequencers. This Technique Has Evolved To Capture Non-Poly(a) Transcripts at Single-Cell Resolution. Today There Are Multiple Flavors of RNA-Sequencing Protocols That Enable the Researcher To Study Not Only Mature Transcripts but Also Immature Transcripts That Are Actively Being Transcribed. Other Methods Such as Crosslinking Followed by Immunoprecipitation Can Also Be Coupled to RNA-Seq To Identify Species of RNA That Are Associated With a Given Protein.

Data analysis of the RNA-seq experiment typically starts by mapping raw reads generated through sequencing to the transcriptomic or genomic reference sequence. Next, the number of reads that are mapped to each annotated transcript is tabulated to compile an expression matrix. The expression matrix is further processed to remove lowly expressed genes or to normalize the gene counts and correct for technical artifacts between samples (Yukselen et al. 2019).

I.4.II. Chromatin Immunoprecipitation Followed by High-Throughput Sequencing: ChIP-Seq

Chromatin Immunoprecipitation (ChIP) is commonly used to detect protein-DNA interactions. In a typical ChIP experiment, cells are crosslinked with

formaldehyde, which allows the proteins and the interacting DNA molecules to be stabilized in the nucleus. Cells are then lysed to extract the crosslinked protein-DNA complexes. The lysate is further sheared either mechanically, enzymatically or both. The sheared lysate is enriched for the protein of interest using an antibody against it. The resulting material is then reverse-crosslinked to extract DNA bound to the protein. This DNA is then ligated to sequencing adaptors and amplified to generate a sequencing library that can be sequenced on any next-generation sequencers. The mechanical and enzymatic shearing used in ChIP-seq experiments is known to have sequence biases (Auerbach et al. 2009). To account for these biases it is necessary to collect control samples. Two types of controls are also typically generated side by side of the samples. In the first type, an IgG mock-antibody is used to pull down the sheared DNA-protein complexes. This pulldown enriches for non-specific DNA-protein complexes which are subsequently processed as the samples and sequenced. The second type of control is “input-DNA”. In this case, no antibody is used for enrichment and a certain amount of DNA-protein complexes are processed as if they were output from enrichment and sequenced.

The analysis pipeline for ChIP-Seq experiments is similar to that for RNA-seq. The sequencing reads generated from the sequencer are aligned to the genome and the signal enrichment for all the genomic locations is then computed by comparing the samples with either IgG or input libraries. This step is called peak-calling. There are several well-established methods for peak calling such as

MACS (Zhang et al. 2008), SPP (Ji et al. 2008) or Scripture (Guttman et al. 2010). Depending on the protein which is pulled down the peaks from a ChIP-seq experiment can be 1) localized, where the binding site is detected as a peak and is typically under 500bp. These peak sizes are expected from transcription factor ChIP-Seq experiments. 2) somewhat broad, where the peaks are around 2kb wide. Such cases are typical for many histone modifications, and 3) extremely broad regions, where the peak widths span several hundred kbs. This type of broad peak is observed in cases such as after pull-down of H3K27me3, a modification observed over large portions of the genome that are repressed.

I.4.III. Chromatin Accessibility

Nucleosome-free regions of the genome have increased accessibility for transcription factors and other components of the transcriptional machinery to bind. These regions are referred to as “Open chromatin” or “nuclease hypersensitive sites”. In 1973 Hewish and Burgoyne showed that active chromatin is preferentially digested by endonucleases (Hewish and Burgoyne 1973) and the resulting digested material showed a periodic banding that captured different levels of nuclear architecture. Since then, endonucleases became a widely used method to create maps of open chromatin regions. After the advent of sequencing technologies, researchers were able to couple the endonuclease digested chromatin material with next-generation sequencing to map these regions genome-wide. The most widely used enzymes for mapping open chromatin regions are DNase-I (DNAse-Seq) (Boyle et al. 2008),

micrococcal nuclease (MNase-Seq) (Hoeijmakers and Bártfai 2018) or a hyper-active Tn5 transposase (ATAC-seq) (Buenrostro et al. 2013). The DNA fragments from these assays display periodic banding patterns. The smaller fragments (<75bp) from such digested material often correspond to regions where TFs are bound. Larger size fragments correspond to mono-nucleosomal fragments (200bp), or di-nucleosomal (around 400bp) and so on. Aggregating all the short fragments genome-wide, from a population of cells, provides a high-resolution map that enables us to detect the footprints of DNA-binding proteins, which are typically between 8-20bps. These footprints are found in all cis-regulatory elements. In theory, such high-resolution chromatin accessibility maps, combined with computational methods to detect TF footprints, can be used in place of generating individual TF binding maps. But this goal has many challenges. First, for a given TF, binding sites across the genome have different affinities, which affect the quality of the footprint. Second, many TFs do not leave a clear footprint, making it difficult for many existing computational pipelines to identify these sites. In Chapter IV, I will discuss a straightforward approach, developed in collaboration with Dr. Nir Yosef's lab, to address these challenges.

I.4.IV. Chromatin Interactions

Chromatin architecture and folding is one of the major mechanisms of regulating gene expression. Chromatin interactions can be studied using two types of methods: microscope-based or molecular biology-based.

I.4.IV.1. Microscope Based Techniques

Previously, light microscopy has enabled us to peek into the chromatin architecture of individual cells at 50-100nm resolution. With this resolution, it was possible to visualize chromosome shapes and their distribution in the nucleus. With the invention of high-resolution microscopes such as electron microscopes, STORM microscopy, etc, the resolution at which the nuclear architecture can be studied has greatly improved. By using fluorescent labeling, we are now able to visualize locations of specific RNA/DNA sequences in the nucleus of live or fixed cells. The fluorescent tags can be inserted into the DNA or delivered to the target location using a dead-cas9 (dCas9) (Ma et al. 2016). The newer methods like oligopaint and oligostrom have revolutionized the microscope-based techniques by making it possible to study chromosome architecture at <= 20nm scale (Beliveau et al. 2017). While these techniques have the capability of measuring the same events within a single cell, they are fundamentally limited by the number of simultaneous events that can be measured, and by the resolution at which they can be measured.

I.4.IV.2. Molecular Biology Based Techniques

Chromosome conformation capture, commonly referred to as 3C, was the first molecular biology-based technique to map chromatin interactions (Dekker et al. 2002). This technique requires the chromatin to be crosslinked and digested with restriction enzymes. The digested chromatin is then ligated, to create chimeric DNA ligation product of interacting regions. The frequency of interaction

between the two regions is then quantified using qPCR probes against them. The major limiting step for this approach is that it depends on knowing the sequence of interacting regions ahead of time. Later, circular chromosome conformation capture (commonly known as 4C) was developed, which can capture all the interacting regions of a given locus (one to all) (Simonis et al. 2006; Zhao et al. 2006). This was followed by chromosome conformation capture carbon copy (5C), where larger genomic regions, and all of its interactions (many to many), could be captured (Dostie et al. 2006). In 2009, the HiC technique was developed to capture all interactions genome-wide (all versus all) (Lieberman-Aiden et al. 2009). In the past few years, many variations of HiC based techniques were developed to increase the resolution and probe enhancer-promoter interactions. These methods include enrichment of the digested chromatin lysate with antibodies of interest (HiChIP) or the use of hybridization probes for sequences of interest (C-HiC) (Mumbach et al. 2016; G. Li et al. 2012; Schoenfelder et al. 2018). All the C-based techniques rely on the same initial steps of crosslinking the chromatin, digesting with enzymes and proximity ligation. However, proximity ligation can only capture pairwise interactions, which does not fully capture all the chromatin interactions that involve more than two interacting genomic regions.

I.4.IV.3. Methods To Capture Higher Order Chromatin Interactions

Recently, three different ligation-free methods were developed to assay higher-order chromosome conformations beyond pairwise interactions: 1) GAM

(genome architecture mapping), 2) ChIA-DROP, and 3) SPRITE (Split pool recognition of interactions by tag extension).

GAM infers chromatin architecture by the presence or absence of contacts in randomly sliced ultra cryosections in a population of nuclei. Like other microscope-based techniques, GAM inherently has single-cell resolution and doesn't depend on digesting the chromatin, but requires specialized equipment and training (Beagrie et al. 2017).

In ChIA-Drop, the nuclear lysate is digested after crosslinking, the lysate is then flown through a microfluidic device such that each chromatin complex is encapsulated in a droplet. The droplet also delivers adapters and barcodes that get ligated to the DNA. All the DNA fragments with the same barcodes are then used to decipher all the contacts that were cross-linked together (Zheng et al. 2019).

SPRITE uniquely barcodes all molecules within a crosslinked complex by repeatedly splitting all complexes across a 96-well plate ("split"), ligating a specific tag sequence onto all molecules within each well ("tag"), and then pooling these complexes into a single well ("pool"). After several rounds of split-pool tagging, each molecule contains a series of ligated tags, which are referred to as a barcode. Because all molecules in a crosslinked complex will travel together through each round of the split-pool-tagging process, molecules in an interacting complex will contain the same barcode, whereas non-interacting molecules in separate complexes will travel independently and receive different

barcodes. The probability that molecules in two independent complexes receive the same barcode decreases exponentially with each additional round of split-pool -tagging. For example, after 6 rounds, there are ~1012 possible unique barcode sequences, which exceeds the number of unique DNA molecules present in the initial sample (Quinodoz et al. 2018).

I.5. Dynamic Transcriptional Regulation

Steady state mRNA expression level is an equilibrium between transcription and degradation rates. Post-transcriptional regulation of mRNAs impacts their stability (Mezan et al. 2013). Thus using a steady state system complicates the understanding of the role of enhancers in gene-expression regulation.

To overcome these limitations I used dynamic cellular systems such as innate immune cells. This system has well defined temporal changes in gene-expression levels which are mostly transcriptionally regulated (Rabani et al. 2011). For example, our lab and others have shown that when macrophages or dendritic cells are stimulated with various pathogen-derived ligands' thousands of genes and cis-regulatory elements undergo changes in well defined temporal clusters (Donnard et al. 2018; Garber et al. 2012; Bornstein et al. 2014; Smale, Tarakhovsky, and Natoli 2014; Link et al. 2018). The genes that are quickest to respond are the ones that do not depend on the synthesis of their activators but rely on the latent transcription factors (Fowler, Sen, and Roy 2011; Escoubet-

Lozach et al. 2011). The second part of the transcriptional cascade is dependent on the proteins encoded by the primary wave.

These robust transcriptional cascades that innate immune cells mount upon stimulation makes them an ideal system to study the contribution of cis-regulatory elements in the transcriptional response. The cascade of events will allow us to study the specific programs in isolation along with global trends.

I.6. Conservation of Regulatory Landscape

Genome-wide maps of histone modifications characteristic of regulatory activity such as H3K27ac, H3K4me1, and H3K4me3, revealed hundreds of thousands of potential regulatory elements in mammalian genomes. It is still unclear what fraction of these elements are actually functional. While experimentally probing each element to determine their activity would be ideal, their sheer number combined with the need to probe in a cell type-specific manner makes systematic experimental approaches unfeasible.

One approach to assess the functional role of regulatory elements is to use comparative genomics, which relies on millions of loss-of-function or gain-of-function experiments selected by nature through evolution. Theoretically, functional elements in the genome should have high constraints across species. By tracking evolutionary changes between species to identify regions with varying conservation we should be able to infer their function. Early comparative studies using DNA sequences alone estimated that about 5% of the human genome is under evolutionary constraint (Lindblad-Toh et al. 2011; Mouse

Genome Sequencing Consortium et al. 2002). Given that less than 2% of the human genome consists of coding genes, a majority of the regions under selection correspond to non-coding regulatory elements or cis-regulatory elements. While the premise of “functional elements having high conservation” is reinforced by the success of comparative sequence analysis in annotating coding sequences, however, the non-coding genomic elements like TF binding sites and enhancer marks undergo rapid turnover across the mammalian phylogeny (Odom et al. 2007; Schmidt et al. 2010; Ballester et al. 2014). This gain or loss of species-specific enhancers across phylogeny is not concomitant with gain or loss of genomic sequence, instead, the majority of species-specific enhancers are composed of ancestral sequences that gain enhancer activity in a species-specific manner (Villar et al. 2015). Given this dichotomy of conserved regulatory elements estimated from sequence conservation and biochemical assays, we hypothesized that enhancers with high evolutionary constraints could be controlling critical gene expression programs, whose misregulation is deleterious.

In chapter II, we used comparative genomics approaches and temporal analysis to understand how different gene expression programs are regulated. As a result, I was able to find that regulatory element conservation is not homogeneous across all enhancers, but rather that it differs across programs. In particular, I found that regulatory elements associated with conserved early induced genes are conserved at twice the rate than those associated with other

expressed genes. Not only regulatory element activity is conserved, but also the underlying sequence is under purifying selection. This allowed us to identify a large set of constrained sequence motifs within active enhancers. This motif lexicon, and their presence in the regulatory elements of response genes, can successfully predict gene induction in both human and mouse dendritic cells.

I.7. Using 3D Interactions To Understand Complex Regulatory Landscape

Two dimensional genome-wide maps of chromatin modifications and accessibility suggest that, on average, a mammalian gene is connected to 5 enhancer regions, and these enhancers can reside at large linear distances from the promoter of that gene (Donnard et al. 2018; González, Setty, and Leslie 2015). Interaction maps created by 4C or promoter capture HiC assays reveal that enhancers do not always regulate the closest gene (measured by linear distance) (Ghavi-Helm et al. 2014), and they can regulate more than one target gene (Fukaya, Lim, and Levine 2016). To further complicate this picture, not only do enhancer-promoter (E-P) interactions vary between cell-types and states (Heintzman et al. 2009), they are also transient (Amano et al. 2009). These pieces of evidence suggest that gene expression is the result of complex multivalent interactions between many different regulatory elements that vary from cell to cell and present a central challenge in mapping E-P interactions.

Beyond linear proximity to promoters, there are many approaches to define enhancers for a given gene. These methods link enhancers to their target

genes based on integrating hundreds of 2D data from diverse cell types or by correlation of signal between enhancers and target promoter or by even pairing enhancers to promoter based on complementary TF binding sites (X. Li and Noll 1994; Merli et al. 1996; Gong and Dean 1993). However, all of these methods result in low to modest prediction accuracy.

Other approaches such as perturbation screens are powerful to identify elements that have regulatory potential. While perturbation platforms have accelerated recently with the successful development of genome engineering based on CRISPR-Cas9 and their adaptation to genome-wide enhancer screens have uncovered critical new insights (Xie et al. 2017; Gasperini et al. 2019), these approaches are limited by the number of enhancers they can validate and further the effect size of many individual enhancers elements are likely to be less than 20% of total gene expression levels, which is below the sensitivity of most current assays.

One alternative is to use the 3D structure of the chromatin to identify functional enhancers in a cell type. Methods such as promoter-capture HiC or HiChIP, have provided the first important insights into these questions by generating a “zoomed-in” view of pairwise E-P interactions. However, they have several intrinsic limitations and, as a result, cannot fully uncover the mechanisms of the complex interactions that regulate gene expression. Specifically, (i) these methods are exclusively pairwise and therefore cannot detect higher-order interactions entailing multiple enhancers interacting with a single promoter, and

(ii) these methods represent an ensemble of interactions that are averaged across a large population of cells, some of which are undergoing active transcription and others that are not.

In chapter III, I used single-molecule chromatin interaction data to differentiate functional from non-functional enhancers within a cell type. I show that using promoter interactions as opposed to linear proximity I can more accurately define functional enhancers. This fine mapping dramatically improved the prediction of transcriptional changes in response to an environmental stimulus. For genes with complex regulatory landscapes, I show two modes of regulatory configuration: stable and transient interactions. By integrating single-molecule chromatin configurations with single-cell RNA-seq data, I showed that genes with transient chromatin interactions have higher variability in gene expression. My data and analysis are the first attempts to map genome-wide combinatorial interactions of enhancers and promoters at single-cell resolution. Our results shed light on the role of chromatin conformations dynamics in driving the transcriptional response. Using this approach, we can start to derive principles, models, and understandings of E-P interactions, to address the fundamental question of how cis-regulatory elements drive cell-type-specific transcription.

I.8. Remodeling of Epigenetic Landscape

Cis-regulatory elements undergo dynamic state changes from being poised to getting activated in response to the extracellular cues and changes in

cellular states. It has been shown previously that H3K4me3 which marks posed promoters is largely invariable across cell types or states while the activity mark H3K27ac (marking transcriptionally active regulatory elements) is specific to cell type or state (Heintzman et al. 2009). The gain or loss of the histone acetylation is mediated by a family of enzymes called histone acetyltransferases (HATs) and histone deacetylases (HDACs). Upon receiving the trigger for change constitutive TFs in the cell undergo post-transcriptional modifications that facilitate them to shuttle into the nucleus, where they bind their DNA binding motif. These modifications that TFs undergo also enable them to interact with other TFs and histone-modifying enzymes (Mokrani et al. 2006; Katto et al. 2013; Gerritsen et al. 1997).

In chapter IV, I will discuss a supervised learning approach developed in collaboration with Yosef lab, to detect chromatin binding events (TF footprints). Using the transcription factor footprints from dynamic and non-dynamic ATAC-seq peaks, we are able to predict the changes in the local H3K27ac signal. By this approach, we prioritized TFs whose binding leads to temporal changes in local chromatin activity.

II. CHAPTER II: Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon

II.1. Preface

This research chapter encompassed work published in Cell Systems, by Elisa Donnard*, Pranitha Vangala*, Shaked Afik*, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, Xiaopeng Zhu, William Diehl, Patrick McDonel, Nir Yosef, Jeremy Luban, Manuel Garber. The publication is entitled "**Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon**" Cell Syst. 2018 Mar 28;6(3):381-394.e7. *co-first authors

II.2. Summary

Most well-characterized enhancers are deeply conserved. In contrast, genome-wide comparative studies of steady-state systems showed that only a small fraction of active enhancers are conserved. To better understand the conservation of enhancer activity we used a comparative genomics approach that integrates temporal expression and epigenetic profiles in an innate immune system. We found that gene expression programs diverge among mildly induced genes while being highly conserved for strongly induced genes. The fraction of conserved enhancers varies greatly across gene expression programs, with induced genes and early response genes, in particular, being regulated by a higher fraction of conserved enhancers. Clustering of conserved accessible DNA sequences within enhancers resulted in over 60 sequence motifs including motifs

for known factors as well as many with unknown function. We further show that the number of instances of these motifs is a strong predictor of the responsiveness of a gene to pathogen detection.

II.3. Introduction

Enhancers act over long chromosomal distances to control gene expression in a cell type-specific fashion (Ong and Corces, 2011). Recent advances in genomic methods have revealed hundreds of thousands of enhancers defined by biochemical signatures that include p300 binding, H3K27ac and H3K4me1 modifications (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Visel et al., 2009). These studies have shown that the vast majority of regulatory elements are species-specific. Furthermore, gain or loss of species-specific enhancers across phylogeny is not concomitant with gain or loss of genomic sequence. Instead, the majority of species-specific enhancers are composed of ancestral sequences that gain enhancer activity in a species-specific manner (Ballester et al., 2014; Kunarso et al., 2010; Mikkelsen et al., 2010; Odom et al., 2007; Schmidt et al., 2010; Villar et al., 2015).

Rapid turnover of species-specific enhancers stands in stark contrast to the highly conserved nature of well-known enhancers that play essential roles in development (Chew et al., 2005; Crocker and Erives, 2008; Lettice et al., 2003), metabolism (Claussnitzer et al., 2015) and viral defense (Panne et al., 2007). Comparative sequence analysis revealed millions of conserved non-coding elements in the human genome that are likely to act as functional enhancers in-

vivo (Pennacchio et al., 2006). Given the general expectation that most functional elements are under purifying selection, there is currently a disconnect between enhancers that are defined by biochemical activity and those defined by evolutionary conservation.

Several arguments have been proposed to reconcile this apparent contradiction between the high turnover rate of biochemical signatures of enhancers observed in comparative studies and the high conservation of a handful of well-characterized examples. One proposed explanation is that typical enhancer elements are redundant, with shadow enhancers that can compensate for the loss of another enhancer (Dunipace et al., 2011; He et al., 2011; Perry et al., 2010). However, redundant enhancers show no relaxation of sequence constraint compared to non-redundant enhancers (Cannavò et al., 2016). Another proposal is that genetic drift may sometimes yield new transcription factor binding sites, eventually leading to novel regulatory elements that make old ones redundant (Ludwig et al., 2000). Accordingly, individual binding sites within enhancers may be shuffled over time and even be replaced by sites occurring on different enhancers. Although both arguments would explain the reduced selective pressure on typical enhancers, they do not explain the apparent strong purifying selection of functionally important enhancers.

An alternative explanation is that most of the biochemically defined enhancers might not be critical in controlling conserved gene regulatory programs. Instead, conserved gene regulatory programs are controlled by a

small subset of conserved enhancers. Here we revisited the question of enhancer conservation by studying the transcriptional regulation of genes that respond to Lipopolysaccharide (LPS). LPS is a cell wall component of gram negative bacteria, that is detected by the TLR4-MD-2 complex (Park et al., 2009). This is a well-defined inducible response in both human and mouse dendritic cells (Amit et al., 2009; Garber et al., 2012; Parnas et al., 2015), which involves hundreds of genes and, in its early stages offers a virtually synchronous response that is mostly transcriptionally controlled (Rabani et al., 2011). Focusing on LPS-responsive genes reduces many confounding factors such as the role of post-transcriptional regulation that make steady state analysis more complex. We focused on the evolutionary profile of enhancers that are associated with both species-specific and shared LPS-responsive genes. Our results reconcile the biochemical and conservation-based definitions of enhancers and demonstrate the importance of evolutionary selection of enhancers in controlling conserved transcriptional programs.

II.4. Results

II.4.I. Transcriptional Dynamics of Human and Mouse DCs in Response to LPS

We generated dendritic cells (DCs) from the bone marrow of two C57BL/6 mice and from human peripheral blood mononuclear cells (PBMCs) from two donors. We stimulated each set of DCs with LPS and collected cells at 0, 1, 2, 4, and 6 hours post-stimulation. We measured genome-wide gene expression by RNA sequencing (RNA-Seq), chromatin accessibility by ATAC-Seq (Buenrostro et

al., 2013) and enhancer activity by chromatin immunoprecipitation of H3K27ac followed by sequencing (ChIP-Seq).

To compare human and mouse response to LPS we focused on genes that could be mapped unambiguously between human and mouse (one-to-one homologs). Immature mouse and human DCs have similar transcriptional profiles with 72% (6,370) of all one-to-one homologous genes detected in at least one species being expressed in both. Among the 3,642 genes that are LPS-responsive in at least one species only 740 have similar expression kinetics (Figure II-1A, STAR Methods). However, induced genes with similar patterns showed greater induction levels (3.7-fold higher on average, Figure II-2A), and were enriched in effectors (cytokines and chemokines $p < 10\text{-}5$, hypergeometric test) and transcription factors (TFs, $p < 0.0001$, hypergeometric test) compared to genes induced in only one species. Overall, the bulk of the differences between mouse and human DCs involve small fold changes and genes that are not critical to the LPS response. There are, however, interesting exceptions of highly induced genes that are species-specific. A well-known example, Nitric Oxide Synthase 2 (NOS2), has an important role in the mouse immune response to microbes but is not induced by LPS in human innate immune cells (Bogdan, 2001; Mestas and Hughes, 2004). Conversely, we find that the T-Cell effector Indoleamine 2,3−dioxygenase (IDO1) gene is highly induced in the human DCs (Mellor and Munn, 2004), but is not induced in mouse DCs.

We next clustered the genes that were responsive in both human and mouse DCs (Figure II-1B, STAR Methods). We observed three broad shared expression trends: genes that were downregulated in both species (clusters D1 and D2), genes that were induced within 1h after LPS stimulation (early induced genes, clusters I1 and I2), and genes that were induced at least 2h after LPS stimulation (clusters I3, I4 and I5). These different clusters showed broad similar expression trends while also reflecting subtle differences in species-specific timing of peak expression. Shared early induced genes were enriched for cytokines and TFs (adjusted $p < 10\text{-}5$, hypergeometric test). Cluster I1 specifically, was 5.4-fold enriched in TFs ($p < 10\text{-}7$, hypergeometric test), including immediate-early genes such as JUN and FOSB. Shared late induced genes included the TFs STAT1 and IRF9 (Figure II-1C), which are involved in autocrine signals from IFNβ and TNFα resulting from LPS detection (Toshchakov et al., 2002).

Although most species-specific genes were induced at relatively low levels, these differences may result from either changes in cis-regulatory elements or from differences in TF expression. We first focused on differences in TF expression. Overall, 530 TFs were expressed in at least one species, of which most (70%) were expressed in both species (Figure II-2B), and most TFs detected only in one species had significant lower expression (Figure II-2C, $p < 10\text{-}15$ Wilcoxon rank-sum test). Further, most TFs that respond to LPS have well conserved kinetics (STAR Methods, Figure II-2D) and although we find specific

TFs having diverging expression patterns, in most cases other members of the same family (defined by TF Class, Wingender et al., 2013) show similar kinetics. For only 15 TFs we found no evidence of compensatory changes, most of these cases involved TFs with a low peak expression or induction (Figure II-2E). These results suggest that TF expression is conserved between mouse and human DCs. Two interesting exceptions are the AP1 factors ATF5 and ATF4, which are highly expressed and induced only in human DCs (Figure II-1D). These two TFs respond to a variety of other stress stimuli, such as amino acid starvation, heat shock and oxidative stress (Harding et al., 2003; Wang et al., 2007a; Watatani et al., 2007), suggesting a human-specific role for cellular stress response in DC response to LPS. We next turned to cis-regulatory elements to further determine the source of changes in expression profiles.

II.4.II. The Epigenetic Landscape of Regulatory Elements in Human and Mouse DC Response to LPS

To define the regulatory landscape of mouse and human DCs we followed a two-step process. First, we mapped candidate enhancer regions using ChIP of histone marks that are typical of transcriptionally active regions (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Shlyueva et al., 2014). We then used ATAC-Seq signal to identify accessible regions within our H3K27ac-defined regions (Buenrostro et al., 2013) (STAR Methods, Figure II-2A).

As in previous studies (Cheng et al., 2014; Vierstra et al., 2014; Villar et al., 2015), we defined Enhancers with Conserved Activity (ECAs) as enhancers

whose sequence could be uniquely mapped across species and which also had H3K27ac signal in both species. We defined Enhancers with SPecies-specific Activity (ESPAs) to include both species-specific sequences with H3K27ac signal and homologous sequences with species-specific H3K27ac signal. Consistent with previous studies (Villar et al., 2015), for the majority of the enhancers and promoters found in one species it was possible to unambiguously identify homologous sequence in the other species (Figure II-3A,B, Figure II-4A and STAR Methods). However, as observed in other systems (Mikkelsen et al., 2010; Schmidt et al., 2010), conservation of H3K27ac signal paints a different picture: While 77% of mouse DC promoters mapped to human sequence with H3K27ac signal, for mouse DC enhancers this fraction is only 25% (Figure II-3B, Figure II-4A). Among transposase-accessible regions within mouse enhancers, only 19% of homologous regions are transposase-accessible in human (Figure II-4B, II-4C). However, among enhancer sequences with conserved H3K27ac signal, 59% also had conserved accessibility in both species. This shows that accessible regions within enhancers and hence TF binding is maintained across evolutionary time whenever the activity of the larger region is also conserved. Overall, the fraction of ECAs (25%) observed in DC enhancers was similar to the one observed between mouse and human liver enhancers (Villar et al., 2015). Thus, in spite of the strong positive selection acting on innate immune cells, the regulatory landscape has not diverged much further than in liver, likely owing to the critical nature of this response for the organism's survival. Since TF

expression is well conserved while cis-regulatory elements have drastically diverged, it appears that most differences in LPS-responsive expression between human and mouse are the result of cis-regulatory changes rather than differences in trans-regulators.

We observed a stronger H3K27ac and ATAC signal in enhancers and promoters that are active in both species, compared to species-specific regions (Figure II-3C, Figure II-4D). This observation could result from a threshold bias to define conserved active loci, with one species having a lower signal that fails to meet the enrichment threshold. However, the H3K27ac signal on the homologous regions of ESPAs was indistinguishable from background (black lines, Figure II-3C, Figure II-4D). Thus, our classification of an active regulatory region as species-specific is not influenced by differing signal intensity.

II.4.III. Enhancers That Are Active in Progenitor Cells Are More Conserved but Are Not Involved in the Response to LPS

Mouse DCs are derived from bone marrow (mBM), whereas human DCs are derived from monocytes. We therefore hypothesized that observed differences in enhancer activity in these cells could be the result of prior activity in progenitor cells. To identify such enhancers we relied on H3K27ac ChIP-Seq data from mBMs (Yue et al., 2014) and generated similar data for human monocytes. Although the fraction of pre-established active enhancers is different in mouse (23% in bone marrow) and human (55% in monocytes), enhancers that are pre-established are more conserved than those that are DC-specific (Figure

II-3D, Figure II-4E). Consequently, pre-established active enhancers are not likely to explain the differences we observed in the transcriptional response to LPS in human and mouse DCs.

The higher degree of conservation among enhancers that are active in progenitors may indicate that they belong to a family of ubiquitous enhancers that have been shown to be more conserved in evolution (Cheng et al., 2014). Consistent with this, nearly half (40%) of the enhancers that are pre-established in mouse bone marrow are also active in liver. Further, we found that pre-established enhancers constitute 39% of all enhancers for genes with rapid downregulation in both species (Cluster D2, Figure II-1B), compared to 23% for all genes. This indicates that ubiquitous enhancers, albeit being more highly conserved than cell type specific enhancers, are not involved in response to stimulus, and are not likely to play an important role in the regulation of LPS response.

II.4.IV. Regulation of Early LPS-Induced Genes Is Both Complex and Conserved

Previous comparative analyses have shown that conserved enhancers are associated with genes involved in specific biological processes (Ballester et al., 2014; Kunarso et al., 2010; Mikkelsen et al., 2010; Schmidt et al., 2010). While there is a slight increase in the fraction of ECAs among shared induced genes compared to enhancers of non-induced or species-specific induced genes, the largest increase (40%, almost double than for non-induced genes) is found on enhancers associated with shared early induced genes ($p < 10\text{-}12$, Fisher exact

test) (Figure II-5A, Figure II-6A). This shows that selection does not act uniformly across all enhancers but rather, that it depends on the particular transcriptional program in which the enhancers function.

Visual inspection of highly induced genes after LPS stimulation such as NFKBIZ, IL6 and PRDM1 (Figure II-3A), suggested that these genes were associated with a high number of enhancers and with super enhancers (Whyte et al., 2013). Such regulatory complexity was previously observed in genes that have a cell type specific regulation during lineage commitment (González et al., 2015). Interestingly, genes with high regulatory complexity (having four or more enhancers) were highly enriched in LPS-responsive genes and particularly, in early induced genes (Figure II-5B, Figure II-6B). Consistent with our initial observation, genes in the top regulatory complexity tier reached higher maximal expression after induction (Figure Figure II-6C). Enhancers that regulate highly induced early genes were also more likely to be conserved. Indeed, on average 2/5 of the enhancers are conserved for shared early response genes with complex regulatory loci, compared to only 1/5 for species-specific early response genes that also have complex regulatory loci (Figure II-5C, Figure II-6D). In general, genes with shared temporal patterns constitute the core of LPS response, and accordingly, their regulation is under strong purifying selection.

II.4.V. Conserved Lexicon Within Accessible Regions

Chromatin accessibility is widely considered critical for transcription factor binding (John et al., 2011; Wang et al., 2012), and we confirmed the strong

preference of TF binding on accessible regions using our previous transcription factor occupancy maps (Garber et al., 2012) (Figure II-9A). As such, DNA accessible regions hold key information related to regulatory activity. Therefore, we next sought to establish the degree to which DNA accessible regions within ECAs are under purifying selection. To this end, we estimated the substitution rate of DNA accessible regions at 10-base resolution (Garber et al., 2009), using a multiple sequence alignment that included 41 mammalian genomes and 2 vertebrate genomes (STAR Methods). Comparison of the substitution rate between DNA accessible regions within ECAs and ESPAs showed a marked reduction in substitution rate (p-value < 10-15, KS-Test, Figure II-7A, S4B). Therefore, ECAs are not only preserved in their activity but there are clear marks of purifying selection in the chromatin accessible sequence within, which is most amenable to TF binding.

To identify sequence elements at the core of ECA function, we clustered conserved 10-mers within ECAs (STAR Methods). Clustering resulted in 66 distinct conserved sequence motifs which we represent by conserved position weight matrices (cPWMs). 31 cPWMs have a clear match to a known transcription factor motif and include all major regulators of TLR4 signaling (STAT, AP1, NFKB, ETV, Figure II-7B, Table II-2). In addition, we identified 35 cPWMs with no clear similarity to any reported motif in public databases (STAR Methods).

Analysis of both known and unidentified cPWMs showed enrichment for genes with specific temporal expression patterns and, in particular, genes with shared response (Figure II-7C, S4C). Importantly, the enrichment of motifs on induced genes was consistent with the expression kinetics of TFs that have affinity for these motifs and recapitulated previous reports (Garber et al., 2012; Medzhitov and Horng, 2009).

To measure the contribution of this conserved lexicon to gene regulation we next trained a random forest classifier to predict if a gene would be strongly induced (> 4-fold) or maintain constant expression following LPS stimulation (STAR Methods). The classifier performed well, achieving a mean area under the curve (AUC) value of 0.75 of the receiver operating characteristic curve (ROC) and a mean AUC value of 0.74 for the precision recall (PR) curve in 10-fold cross-validation (Figure II-7D). This confirms the ability of cPWMs to predict gene induction, but also suggests that cPWM instances alone are not sufficient predictors.

Importantly, when we applied the model we trained in mouse to predict expression induction in human, it performed with similar accuracy and precision, achieving an AUROC of 0.68 and an AUC value of 0.63 for the PR curve (Figure II-7D). Motifs of the key regulators such as NFKB, AP1, STAT and EGR along with several novel GC rich motifs are amongst the top classifying features (Figure II-7E).

II.4.VI. Enhanceosomes in Conserved Innate Immune Responses

Enhancers are thought to function in two broadly different mechanisms (Arnosti and Kulkarni, 2005). In enhanceosomes, TFs act cooperatively and their binding results in an on/off signal, where loss of even one TF binding site profoundly disrupts the function of the enhanceosome. Billboards on the other hand, are modular enhancers where the binding of each TF is not necessary for enhancer activity but rather has an additive or synergistic effect.

The prototypical enhanceosome is the IFNβ proximal enhancer, which requires the assembly of 6 TFs to induce IFNβ expression (Thanos and Maniatis, 1995). Mutations that disrupt a single binding site disrupt the enhancer and are highly deleterious. Consistent with this, the IFNβ enhanceosome sequence is more highly constrained than the protein coding sequence of IFNβ, the gene it regulates (Figure II-8). Since the effect of mutations on enhanceosomes can be highly penetrant, we sought to identify and catalog enhancers that have characteristics typical of enhanceosomes and that may help prioritize non-coding mutations associated with immune disease.

We scanned for candidate enhanceosome regions in chromatin accessible regions within ECAs that were 1) Bound by at least six TFs, based on our previous binding maps of 14 TFs and 2) Had a large portion (> 30%) of their sequence conserved. Our scan identified 80 chromatin accessible regions (Figure II-10 for example & Table II-3) that resemble enhanceosomes, such as the IFNβ proximal enhancer (Figure II-8). Consistent with their innate immune

specific function, genes associated with these conserved, highly bound regions tend to have similar temporal induction in both human and mouse ($p < 0.01$ Fisher's exact test) and are highly enriched in IRF1, RELA (also known as p65) and RUNX1 binding ($p < 10\text{-}10$, Fisher's exact test). The high evolutionary sequence constraint that we required to define enhanceosome candidates translates to low variation across the human population. Indeed, human regulatory regions with similar evolutionary constraint are depleted of Single Nucleotide Polymorphisms (SNPs) having an average of only 25 SNPs compared to an average of 400 (and a minimum of 369) in similarly sized genomic regions.

II.4.VII. Regulatory Regions With Conserved Activity and Temporal Patterns Regulate Highly Induced Genes With Shared Kinetics

Response to LPS affects both the acetylation and chromatin accessibility of thousands of enhancers (Figure II-11A, S6A-C). Although the chromatin state of most enhancers (72%) is unaffected by LPS, enhancers that show temporal kinetics tend to associate with genes having similar transcriptional kinetics. Indeed, regions whose DNA accessibility increases upon LPS stimulation are associated with induced genes (1.6-fold enrichment) while regions that close over time are associated with downregulated genes (2-fold enrichment, Figure II-11B). We further observed a clear enrichment of cPWMs, including NFKB, STAT and AP1 motifs, on DNA accessible regions that show increased ATAC signal after LPS stimulation. On the other hand, cPWMs associated with ETV

and STAT transcription factor families are enriched in accessible DNA regions that become less accessible in response to LPS. Enrichment of ETV and STAT motifs on regions that lose availability is consistent with their reported repressive function (Icardi et al., 2012; Mavrothalassitis and Ghysdael, 2000) (Figure II-11C). It is interesting that STAT motifs are enriched in both down and upregulated elements. These motifs may recruit different members of the STAT family or attract complexes involving different TFs that modulate the STAT TF function. Our previously generated mouse binding data for STAT1 and STAT2 shows that these proteins bind mostly to regions that become increasingly accessible upon LPS stimulation. This suggests that motifs in regions whose DNA availability decreases after LPS stimulation are likely bound by different STAT TFs or other factors that can bind this motif.

To further determine the importance of cPWMs in regulating the LPS response, we proceeded to build a random forest classifier as above, but this time we associated each cPWM with three features per gene: the number of cPWMs in regulatory regions with increased, diminished or unchanged DNA accessibility upon LPS stimulation. This dramatically improved the model performance which now showed an average AUROC of 0.82 in mouse in a 10-fold cross-validation and an AUROC of 0.78 when applied to human (Figure II-11D). This highlights the importance of the chromatin context and helps explain the weaker performance of a model that was trained on sequence alone.

Given that regions with LPS-responsive chromatin dynamics were important when evaluating sequence features, we next investigated the conservation of DNA accessibility dynamics. Interestingly, although regions with LPS-induced DNA accessibility are present in both human and mouse (28% and 30%, respectively), very few are LPS-responsive in both. By simultaneously clustering ATAC-Seq peaks from ECAs that had significant LPS-induced signal changes in at least one species (Figure II-12D), we found that only 500 such regions (13%) are responsive in both mouse and human DCs (Figure II-11E).

These 500 regions are associated with 325 genes, of which 57% have similar expression kinetics in human and mouse, while only 21% of all the expressed genes have similar expression patterns in both species ($p < 10^{-20}$, Fisher-exact test, Figure II-11F). Genes associated with these regions have much higher induction levels and reached higher maximal expression than other genes with no difference in baseline expression (Figure II-11G-I, 6G: $p < 2.2e^{-16}$ Wilcox-rank test, 6H: $p < 2.2e^{-16}$ Wilcox-rank test, 6I: not significant Fisher-exact test). They include cytokines (e.g IL1B, IL6) and key transcription factors (e.g. REL, NFKB1, BCL2, NFKBIZ) (Figure II-11J, p-adjusted < 0.004). Regions with conserved dynamics are enriched near genes with similar temporal dynamics and have maintained enhancer activity since the rodent/primate divergence. This suggests that they are crucial elements regulating this set of genes.

II.4.VIII. Transposable Elements Are Enriched in Cis-Regulatory Regions of LPS-Induced Genes

Most cis-regulatory elements are composed of ancestral sequence (Cheng et al., 2014; Villar et al., 2015) (Figure II-3B). Therefore turnover of ancestral activity rather than sequence seems to be the major force reshaping regulatory regions. Sequence changes can still be an important source of difference between the human and mouse response. Since lineage specific transposable elements (TEs) have been shown to significantly modify transcriptional networks (Lowe and Haussler, 2012; Wang et al., 2007b), we next sought to determine whether TEs have contributed to regulatory sequence involved in the LPS response. We identified 25 families of TEs in mouse and 15 in human that are enriched in regulatory regions (enhancers or promoters) of induced genes (Figure II-13A). These enriched TE families fall into two categories: those that were actively mobile prior to the human-mouse divergence, and newer elements that have only been active in either the mouse or human lineage. The majority belong to one of the ancestral TE families of Mammalian-wide interspersed repeats (MIRs), with MIR3 elements being the most enriched (Figure II-13B) and having the largest number of elements within regulatory regions. MIR elements are some of the oldest (Smit and Riggs, 1995) and most conserved families of mobile elements (Jjingo et al., 2014), and have been reported to contribute to the regulation of cell type specific expression (Jjingo et al., 2014). Our data further suggests that MIRs, and MIR3 in particular,

have been co-opted into regulation of innate immune responses prior to the euarchontoglires ancestor. As one might expect for important regulatory sequences, we observed that MIRs have been under clear purifying selection (Figure II-14A).

Lineage specific TEs enriched in DC regulatory regions include mainly endogenous retroviral Long Terminal Repeat (LTR) elements. We found that elements from these families (ORR1E in mouse and THE1A and THE1C in human) tend to be positioned at the most accessible regions within enhancers, possibly indicating a role in creating or facilitating opening of chromatin that is more favorable to transcription factor binding and more likely to function as a regulatory element (Figure II-13C).

II.5. Discussion

Massive parallel sequencing has revealed hundreds of thousands of active non-coding regions, most of which are classified by their chromatin signatures as enhancers or long noncoding RNAs (lncRNAs). Comparative analyses of enhancers and lncRNAs have shown that although the majority are encoded by ancestral sequence, their activity is generally species-specific (Chen et al., 2016; Cheng et al., 2014; Kutter et al., 2012; Necsulea et al., 2014; Ponjavic et al., 2007; Ulitsky, 2016; Vierstra et al., 2014; Villar et al., 2015; Washietl et al., 2014). Here we showed that a higher fraction of enhancers that regulate specific pathways tend to be conserved over longer evolutionary time.

As opposed to previous studies, we used a dynamic system and focused on temporal expression patterns rather than steady state expression. In this system, changes in mRNA levels in early time points are mostly the result of transcription rather than post-transcriptional processes (Rabani et al., 2011); this helps isolating and measuring the contribution of cis-regulatory elements to expression changes. Temporal analysis also allowed us to study different regulatory programs individually rather than analyzing all regulatory programs together or by broad functional classes (Figure II-1B). As a result, we were able to find that regulatory element conservation is not homogeneous across all enhancers, but rather that it differs across programs. We find that regulatory elements associated with shared early-induced genes are conserved at twice the rate than those associated with other expressed genes. Not only regulatory element activity is conserved, but also the underlying sequence is under purifying selection. This allowed us to use comparative sequence analysis to identify a large set of constrained sequence motifs within active enhancers. Functional validation of these enhancers as well as the novel motifs we found will be critical, but this study provides a clear path towards the goal of functionally characterizing a well-defined set of regulatory regions involved in well-understood cellular processes.

It is interesting that, besides enhancers associated with shared induced genes, the other set of enhancers preserved since the euarchontoglires ancestor are ubiquitous or active in progenitor cells but are not associated with genes

induced by TLR4 signaling. Instead, these enhancers tend to lose active marks following LPS stimulation. This is consistent with previous observations that basic cellular processes are passively downregulated upon induction of a large transcriptional program (Cheng et al., 2009; Garber et al., 2012), perhaps due to a shift of limited resources towards the response to immune challenge.

The greater conservation of enhancers associated with early induced genes is surprising, with conserved enhancers accounting for 40% of all enhancers associated with these genes. This arises an interesting question: why are the regulatory elements of early-induced genes under stronger selection? It is reasonable to argue that this initial wave of transcription triggers a program that, although necessary for immune defense, is deleterious to the individual when misregulated. Tight control of the initiation of the program may be critical to avoid unwanted harm. It is also interesting that in our previous analysis of mouse DC enhancers we observed a low degree of sequence constraint of most enhancers, and concluded that early-induced genes were regulated by a highly redundant regulatory architecture that functioned by recruiting many different TFs in a nonspecific fashion. Our comparison with human DCs paints a more nuanced picture. Early induced genes are regulated by a mix of highly constrained enhancers that have been preserved over hundreds of millions of years and newly evolved species-specific enhancers. The ECAs have clear signatures of undergoing purifying selection and may be necessary for induction. Nonetheless, the majority of enhancers is species-specific and may play redundant, subtler

roles or have no impact on gene expression. Further functional studies will be needed to determine how different enhancers function and how they interact to produce reproducible, precise patterns of expression.

Our study sheds some light on the long-standing question of how selection acts on gene expression (Gilad et al., 2006). Although our study was not designed to answer this, we find two very clear modes of selection. On one hand, highly induced genes tend to have shared induction and are regulated by conserved regulatory elements. These observations are consistent with strong stabilizing selection. On the other hand, there is great divergence among genes with mild induction, which is consistent with neutral selection (Gilad et al., 2006). We reason that, while mutations that disrupt the level and timing of highly induced genes may have strong deleterious effect, for genes that are mildly induced, changes are tolerated.

Our comparative map provides a unique resource for future studies of *in-vitro* derived DCs. It provides a reference map of the genomic elements that can be mapped and translated from a mouse model to human biology. Further, recent reports on underlying differences in the cell types obtained in mouse and human DC in-vitro cultures (Helft et al., 2015) highlights the need to compare these two systems at the molecular level. In this work, we focused on understanding both the similarities and differences between the two. Given the overall similarity in TF expression, this system offers a deep platform to understand the impact of *cis*-regulatory changes on expression.

II.6. Methods

II.6.I. Key Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
List of 147 data-sets used in this study	This paper	Table II-4
Human 10-mers substitution rates	This paper	http://garberlab.umassmed.edu/data/conservation/hg19/omega/
Mouse 10-mers substitution rates	This paper	http://garberlab.umassmed.edu/data/conservation/mm10/mm10.omega
Software or Algorithms		
gkm-SVM	(Ghandi et al., 2016)	v1.3
Spectral clustering	This paper	https://github.com/nimezhu/ClsViz
Trimmomatic	(Bolger et al., 2014)	V0.32
Bowtie2	(Langmead and Salzberg, 2012)	v2.2.23
Samtools	(Li et al., 2009)	v0.1.19
DESeq2	(Love et al., 2014)	v1.10.1

Bedtools	(Quinlan and Hall, 2010)	V2.25.0
MACS2	(Zhang et al., 2008)	V2
IGVtools	(Robinson et al., 2011)	V2.3.31
RSEM	(Li and Dewey, 2011)	v1.2.28
SiPhy	(Garber et al., 2009)	https://github.com/garber-lab/siphy
Antibodies & Reagents		
H3K27ac	Diagenode	C15410196
H3K4me3	EMD Millipore	05-745R
Ovation Human FFPE RNA-Seq Library System	NuGen	340
Ovation mouse RNA-Seq Library System	NuGen	348
RNeasy mini plus kit	Qiagen	74134
Nextra TDE-1 transposase,	Illumina	FC-121-1030
Covaris tru-ChIP Chromatin Shearing and Reagent Kit	Covaris	520154
Agencourt AMPure XP	Beckman Coulter	A63880

GMCSF	Miltenyi	130-095-735

II.7. CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Manuel Garber (Manuel.Garber@umassmed.edu).

II.8. EXPERIMENTAL MODEL AND SUBJECT DETAILS

II.8.I. Human Subjects:

Anonymous, healthy donor leukopaks (New York Biologics, Southampton, NY), were used in accordance with UMMS-IRB protocol ID #H00004971

II.8.II. Mice:

All mice were housed in specific pathogen-free condition in accordance with the Institutional Animal Care and Use Committee of the University of Massachusetts Medical School. C57BL6 female mice were euthanized at 6-8 weeks of age to harvest bone marrow.

II.9. METHOD DETAILS

II.10. Cell Culture

All cells were maintained at 37° C in 5% CO2 humidified incubators.

II.10.I. Human Monocyte-Derived Dendritic Cells

Human dendritic cells were derived from peripheral blood mononuclear cells (PBMCs) isolated from de-identified, healthy donor leukopaks (New York Biologics, Southampton, NY), in accordance with UMMS-IRB protocol ID

#H00004971. Mononuclear leukocytes were isolated by gradient centrifugation on Histopaque-1077 (Sigma-Aldrich, St. Louis, MO). CD14+ mononuclear cells were enriched via positive selection using anti-CD14 antibody MicroBead conjugates (Miltenyi, San Diego, CA), according to the manufacturer's protocol. CD14+ cells were then plated at a density of 1 to 2 x 106 cells/ml in RPMI-1640 supplemented with 5% heat-inactivated human AB+ serum (Omega Scientific, Tarzana, CA), 20 mM L-glutamine (ThermoFisher, Waltham, MA), 25 mM HEPES pH 7.2 (Sigma-Aldrich), 1 mM sodium pyruvate (ThermoFisher), and 1 x MEM non-essential amino acids (ThermoFisher). Differentiation of the CD14+ monocytes into dendritic cells (human DCs) was promoted by addition of recombinant human GM-CSF and human IL-4; cytokines were produced from HEK293 cells stably transduced with pAIP-hGMCSF-co or pAIP-hIL4-co, respectively, as previously described (Reinhard et al., 2014), with each cytokine supernatant added at a dilution of 1:100.

II.10.II. Mouse Bone-Marrow Derived Dendritic Cells

Mouse dendritic cells were derived from bone marrow harvested from 6-8 week old female C57BL6 mice. Bone marrow was then dissociated into single cells and filtered through 70um cell strainer. The cells were then incubated with red blood cell lysis buffer for 5 minutes. To differentiate bone marrow to dendritic cells, bone marrow cells were plated at 200,000 cells/mL in non-tissue culture treated plates. These cells were supplemented with media on day 2 and day 7. On day 5 cells were harvested and resuspended in fresh media. On day 8 all the

floating cells were collected as mouse bone marrow derived dendritic cells. The media used for culturing and differentiating contains RPMI (Gibco) supplemented with 10% heat inactivated FBS (Gibco), β-mercaptoethanol (50uM, Gibco), MEM non-essential amino acids (1X, Gibco), sodium pyruvate (1mM, Gibco), and GM-CSF (20 ng/ml; Miltenyi).

II.11. Library Preparation and Sequencing

II.11.I. RNA-Seq

Total RNA was isolated from frozen dendritic cell pellets using the RNeasy mini plus kit (QIAGEN). The RNAs were additionally treated with RNase-free DNase I for 15 minutes at room temperature to eliminate most genomic DNA. RNA-Seq libraries were prepared from 70 ng of starting RNA using the Ovation Human FFPE RNA-Seq Library System (NuGEN) or Ovation mouse RNA-Seq Library System (NuGEN), according to the manufacturer's protocol. Fragmentation of the cDNA was achieved by sonication using the M220 sonicator (Covaris) with the following conditions: sonication time = 350 seconds; temp = 20°C; peak power = 50; duty factor = 20; cycles/burst = 200. The quality of the isolated RNA, as well as of the final libraries, was assessed using the 2100 Bioanalyzer (Agilent) and Qubit (Invitrogen). The libraries were pooled according to donor in equimolar ratios and denatured. Pooled libraries were sequenced for 2 x 100 cycles to obtain paired-end reads, using a HiSeq 2000 (Illumina) for human DCs and 2 x 75 cycles using Nextseq 500 for mouse DCs.

II.11.II. ATAC-Seq

For each time point, 5 x 10^5 scraped DC's were collected by centrifugation 500 x *g* for 5 min. and lysed for ATAC-seq following the protocol described in (Buenrostro et al., 2015). Each sample was tagmented using 12.5 ul Nextera TDE-1 transposase (Illumina) for 30 minutes at 37, then quenched by addition of 5 volumes DNA Binding Buffer (Zymo Research) and cleaned using Zymo Research DNA Clean and Concentrator-5 columns according to the supplied protocol. Tagmented DNA was PCR-amplified using indexed primers as described in (Buenrostro et al., 2015), using total cycle numbers for enrichment as determined empirically by qPCR to minimize PCR duplicates. The resulting libraries were purified twice by Zymo Research DNA Clean and Concentrator-5 columns using a ratio of 5:1 DNA Binding Buffer:Sample, and quantified by Qubit HS-DNA Assay (Thermo Fisher Scientific) and Bioanalyzer High-Sensitivity DNA (Agilent Technologies). Final ATAC-seq libraries were pooled (equimolar) and sequenced on an Illumina Nextseq 500.

II.11.III. ChIP-Seq

Harvest and Formaldehyde crosslinking. For each timepoint and donor, 5-7 x 10^6 unstimulated or LPS-stimulated hDCs were harvested by scraping in medium and centrifugation at 500 x *g* for 5 minutes. Each cell pellet was washed once with 2 mL PBS and gentle flicking of the tube, followed by centrifugation at 500 x *g* for 5 min. Cells were uniformly resuspended in 1 mL 1X Fixing Buffer A from the Covaris tru-ChIP Chromatin Shearing and Reagent Kit and fixed by

adding 1 mL 2% methanol-free formaldehyde (Thermo Fisher Scientific) diluted in 1X Fixing Buffer A (1% formaldehyde final, 2.5-3.5x10^6 cells/mL) and rotated end-over end for 5 min. at room temperature. Fixation was quenched by adding 240 mL Quenching Buffer E (Covaris tru-ChIP kit) and rotating for an additional 5 min. Purified BSA was then added to 0.5% w/v final to prevent cell adherence to the tube, and crosslinked cells were harvested by centrifugation, 500 x *g* for 5 min. at 4°C. Crosslinked cells were washed twice in 2 mL ice-cold PBS + 0.5% BSA with centrifugation as above, and aliquoted evenly into 3 fresh 1.5 mL tubes during the second wash. Cells were finally pelleted by centrifugation at 16,000 x *g*, flash-frozen as dry pellets in liquid nitrogen, and stored at -80°C.

Lysis, Shearing, and Quantification. Individual crosslinked cell pellets (1.5-2 x 10^6 cells each) were lysed according to the Covaris tru-ChIP Chromatin Shearing and Reagent Kit instructions. Following lysis, nuclei were resuspended in 130 mL ice-cold Shearing Buffer D3 and transferred to 1.5 mL BioRupter Pico Microtubes (Diagenode) on ice. Chromatin was sheared to uniform fragment lengths (150-400 bp) by sonication at 4°C in a BioRupter Pico (Diagenode) set to 6 cycles of 30s ON and 30s OFF. Sheared chromatin was diluted in 10 volumes of ChRIPA buffer (1X PBS, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5% sodium deoxycholate, 1% Igepal CA-630, 0.1% SDS, 1X Roche cOmplete Protease Inhibitor Cocktail) and insoluble material was removed by centrifugation >15,000 x *g* for 10 minutes. Lysate was pre-cleared against 60 mL Dynabeads Protein A (Thermo Fisher Scientific) per 10^6 cells for 2h at 4°C with end-over-end

rotation followed by two rounds of magnetic bead removal and transfer to fresh tubes. 2% of pre-cleared lysate was removed for DNA quantification and the remaining lysate was either flash-frozen in liquid nitrogen and stored at -80◦C, or stored overnight at 4◦C for use in immunoprecipitation. For quantification, 2% pre-cleared lysate was treated with 10 mg RNase A (Thermo Fisher Scientific) for 30 min. at 37◦C, followed by addition of 100 mg Proteinase K (New England Biolabs) and crosslink reversal overnight at 65◦C. DNA was purified using DNA Clean and Concentrator-5 columns (Zymo Research). Average sheared DNA fragment sizes were determined by agarose gel and chromatin yield was estimated by Qubit HS-DNA Assay. 50-100 ng purified DNA was saved as Input.

Chromatin Immunoprecipitation. Antibodies used for ChIP were rabbit anti-H3K27ac (Diagenode C15410196) and rabbit anti-H3K4me3 (EMD Millipore 05-745R). 1 mg antibody was added to 0.5 mg (anti-H3K27ac) or 1 mg (anti-H3K4me3) pre-cleared crosslinked lysate and incubated overnight with continuous mixing at 4◦C. IgG/chromatin complexes were captured for 1h at room temperature on 25 mL Dynabeads Protein A that were pre-blocked for at least 1h with Blocking Buffer (1X PBS, 0.5% BSA, 0.5% Tween-20). Complexed beads were washed 5 times with ice-cold ChRIPA Buffer, twice with room temperature RIPA-500 Buffer (10 mM Tris pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS), twice with ice-cold LiCl Wash Buffer (10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal CA-630, 0.5% sodium deoxycholate), and twice with ice-cold TE buffer. Each chromatin sample

was eluted from beads using 50 ul Direct Elution Buffer (10 mM Tris pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.5% SDS) and supplemented with 20 mg RNase A, incubating for 30 min. at 37◦C. 20 mg glycogen was added to each bead/eluate suspension, and crosslinks were reversed by addition of of 50 mg Proteinase K and incubation at 37◦C for an additional 2h, followed by overnight at 65◦C. Dynabeads were removed by magnet capture, and the supernatant was mixed thoroughly with 2.3 volumes of Agencourt AMPure XP (Beckman Coulter) bead suspension and incubated for 10 minutes at room temperature prior to bead capture and washing. Purified DNA was eluted in 10 mM Tris pH 8.0.

Library Preparation and Sequencing. Sequencing libraries were prepared from half of each ChIP sample and 50 ng Input DNA using the Ovation Ultralow System V2 kit (NuGEN) according to supplier's instructions, with the total numbers of enrichment PCR cycles determined empirically for each sample by qPCR to minimize PCR duplication rates. Barcoded libraries were quantified using Qubit HS-DNA Assay, qualified using Agilent Bioanalyzer High-Sensitivity DNA, and pooled for sequencing on Illumina Nextseq 500.

II.12. Quantification and Statistical Analysis

II.12.I.1. Alignment and processing of reads

RNA-Seq: *Trimmomatic-0.32* (Bolger et al., 2014) was used to remove 5' or 3' stretches of bases having an average quality of less than 20 in a window size of 10. Only reads longer than 36 bases were kept for further analysis. Reads were then aligned to human or mouse ribosomal RNA using *Bowtie2* v2.2.3

(Langmead and Salzberg, 2012) with parameters *-p 2 -N 1 --no-unal*. All reads mapped to rRNA were discarded from further analysis. *RSEM* v1.2.28 (Li and Dewey, 2011) was used to estimate gene expression in Transcripts per Million (TPM), with parameters *-p 4 --bowtie-e 70 --bowtie-chunkmbs 100 --strand-specific*. RSEM is configured to use *Bowtie* v0.12.9. Quantification was run against the transcriptome (RefSeq v69 downloaded from UCSC Table Browser (Pruitt et al., 2012). Genes with more than 10 TPM in any time point were considered expressed, and genes that did not achieve this threshold were removed from further analysis. Moderate batch effects were observed between samples from different mice and between the two human donors. We used the log transformed TPM normalized expression values as input to *ComBat* (package *sva* version 3.18.0) (Johnson et al., 2007; Leek et al., 2012) with default parameters and a model that specified different donors or mice as batches. Corrected TPM values were transformed back to read counts using the expected size of each transcript informed by *RSEM*. We only considered genes with at least 10 TPMs in at least one replicate at any time point.

ATAC-Seq: Paired-end reads were trimmed to remove adapter sequence using *Cutadapt* version 1.3, and then aligned with *Bowtie2*, version 2.1.0, parameter *–X 2000*. Reference genome hg19 was used for human samples and mm10 for mouse samples. The alignments were then filtered using *Samtools (Li et al., 2009)*, version 0.0.19, to remove (i) PCR duplicates, as identified by Picard's *MarkDuplicates*, and (ii) aligned reads with mapping quality below 4.

While the reads were aligned as paired-end to optimize the alignment accuracy, the alignments were then further processed as if they were aligned single-end sequence data, so that each aligned read corresponded to a Tn5 cut-site.

Peak Calling: Each aligned read was first trimmed to the 9-bases at the 5'-end, the region where the Tn5 transposase cuts the DNA, and then extended 10-bases upstream and down, for smoothing. Peaks were called using these adjusted 29-base aligned reads with *MACS2 (Zhang et al., 2008)*], parameters *--bw 29 --tsize 29 and --qvalue 0.0001*. For visualization, the adjusted aligned reads were converted to tdf files using *IGVTools*, version 2.3.31 (Robinson et al., 2011) (*IGVtools count –w 5*).

Quality Control: Following the standard practice (Buenrostro et al., 2015), for each sample, we examined the fragment length distribution, as well as a comparison of the aggregate nucleosome signal to the aggregate nucleosome-free signal over transcription start sites for those genes found to be expressed for at least one time point in our RNA-Seq time series. Signal-to-noise ratios were computed for the peaks as f/(1 –f) where f is the fraction of reads overlapping peaks.

ChIP-Seq: Along with in house generated data we also analyzed publicly available data for mouse bone-marrow progenitors generated by the Encode consortium (Accession: GSM1000108). Paired-end reads were trimmed to remove sequencing adapters and leading and trailing bases with quality scores less than 5. Reads that were longer than 36 bases after trimming were kept for

further analysis. The reads were then aligned to human reference genome hg19 or mouse genome mm10 using *Bowtie2* with options *-k 1 --un-conc* to filter out reads that map to multiple locations in the genome and that align un-concordantly. Duplicated reads were filtered out using picard-tools-1.131 *MarkDuplicates* function. Peaks were then called using *MACS2* with *--bw=230 --tsize=75 and --qvalue 0.0001*. Alignment files were also converted to tdf format using *IGVtools count* function using *-w 5 --pairs options* for visualizing. H3K27ac ChIP-Seq peaks were filtered to retain only the peaks that are two-fold enriched over input.

II.13. Gene Classification and Clustering

Homologs: All our analysis were restricted to genes that had homologous pair between human and mouse defined in the Homologene release 68 (NCBI Resource Coordinators, 2016), resulting in a list of 16,500 one to one homologous gene pairs.

Gene Classification: The expressed gene list was filtered to include only genes with homologs as defined by the previous step. We used the batch corrected (see above) counts per gene to identify differentially expressed genes by at least 2 fold between unstimulated cells (time 0) and any time point following LPS stimulation whose change in expression was significant (p-adjusted < 0.05) according to the package *DESeq2* (v1.10.1) (Love et al., 2014) in R (v3.3.1). Due to the large transcriptional changes observed in this system, we turned off the fold change shrinkage in *DESeq2* with *betaPrior=FALSE* and we added a

pseudocount of 32 to all timepoints to avoid spurious large fold change estimates from lowly abundant genes. Genes were then classified based on their response to LPS stimulation in each species (induced, downregulated or non-responsive).

Clustering expression patterns: For genes expressed in both species and presenting similar response following LPS stimulation (induced in both species or downregulated in both), we applied a spectral clustering approach (von Luxburg, 2007) to identify genes with conserved expression patterns in mouse and human. Briefly, let {g1,g2,g3,…,gn} represent the set of response genes, and let E_{Mi} and E_{Hi},$1 \leq i \leq n$, represent the expression time courses in TPM for gene gi in mouse and human respectively. Further, let $\rho_M = [\rho_{Mij}]$, $1< i,j< n$ represent the Pearson correlation coefficient matrix, where ρ_{Mij} is the coefficient of correlation of E_{Mi} with E_{Mj}. The human correlation coefficient matrix, ρ_H is defined similarly. We define similarity matrices $[s_{Mij}]$ and $[s_{Hij}]$, for mouse and human respectively, where $s_{Mij} = \exp(-(\sin(\cos^{-1}(\rho_{Mij})/2)^2)$, and $s_{Hij} = \exp(-(\sin(\cos^{-1}(\rho_{Hij})/2)^2)$. Then the matrix $W = [w_{ij}] = [s_{Mij}s_{Hij}]$ defines a similarity matrix for {g1, g2, … , gn} and can be viewed as an adjacency matrix for a weighted graph, where each gene represents a node in the graph. We associate to W its graph Laplacian L = D - W, where D is the diagonal degree matrix with entries $d_{ii} = \sum_{j=1}^{n} w_{ij}$. L is positive, semi-definite and therefore has n real non-negative eigenvalues, λ_i, $1 \leq i \leq n$, which we list in descending order, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. We select k, the number of clusters, to be the smallest positive integer such that $(\lambda_1 + \lambda_2 + \ldots + \lambda_k)/tr(L) > 0.95$, where tr(L) is the trace of L. We then construct a matrix with columns set to the first k

eigenvalues of L and apply k-means clustering to the rows of this matrix to cluster the genes into k distinct clusters. The python script used for spectral clustering is available on https://github.com/nimezhu/ClsViz.

We analyzed enrichments for specific Gene Ontology categories using *clusterProfiler* (Yu et al., 2012).

II.14. Transcription Factor Network

We sought to first determined the extent to which the TF network in response to LPS is conserved between human and mouse DCs. To systematically explore core changes in the regulatory network, we compared the overall trends of the 258 transcription factors that responded to LPS-stimulation in at least one of the two species (Figure II-2D). We calculated the Pearson correlation between the expression patterns across all timepoints for TFs with response to LPS per species. The resulting distance matrix was hierarchically clustered and displayed as a heat map. We chose the number of groups in each clustering by visual inspection of the dendrogram and selection of a threshold. Membership in each cluster was then compared across species to identify the corresponding groups.

Transcription Factor Network Overview: There are 3 large co-regulated groups of transcription factors with no major changes between the species, and a fourth cluster in mouse composed of only 8 TFs (Table II-1) with very small changes in expression in mouse (< 2 fold), that are scattered across all three human clusters. The largest cluster in mouse contained 115 genes that were

downregulated following LPS treatment. Further, 73% of the factors that were also expressed in human remained in the same cluster and showed a similar transcriptional downregulation pattern in human (Figure II-2D, top right). Similarly, the vast majority (77%) of induced transcription factors were induced in both species, with 17 factors (19%) having different induction timing in each species (Table II-1). The largest of the induced clusters (pink cluster, Figure II-2D), contained mostly TFs with conserved kinetics (66% in mouse and 57% in human, Figure II-2D, bottom right). This group included members of the NFKB, IRF, and STAT families (Figure II-1C). The smaller cluster of induced transcription factors also contained important rapidly upregulated TFs (blue cluster, Figure II-2D, middle right), including members of the FOS and JUN families, as well as MAFF, PRDM1, and EGR3, all of which show a conserved pattern in the human response. 17 mouse-specific and 12 human-specific TFs were induced by LPS. Interestingly, to the best of our knowledge, none of the species-specific factors have been studied in the context of innate immune signaling. Two mouse-specific TFs, ID1 and SIX1, are highly induced in mouse, although not detectable in human. Similarly, MSC is highly induced in human DCs but has no detectable expression in mouse DCs. Outliers such as these however, are rare, and most TFs with different responses in mouse and human DCs have moderate induction compared to genes with conserved response.

II.15. Substitution Rate Scan

We used SiPhy (Garber et al., 2009) to compute the substitution rate () for every 10-mer in the mouse and human genomes. For human we used the vertebrate multiple sequence alignment available from the UCSC genome browser for the hg19 assembly. We removed the vertebrates danRer6, petMar1, oryLat2, gasAcu1, fr2, tetNig2 which left us with the following phylogeny: ((((((((((((((((hg19:0.006653,panTro2:0.006688):0.002482,gorGor1:0.008783):0.009697,ponAbe2:0.018183):0.040003,rheMac2:0.008812):0.002489,papHam1:0.008723):0.045139,calJac1:0.066437):0.057049,tarSyr1:0.137822):0.010992,(micMur1:0.092888,otoGar1:0.1295):0.035423):0.015348,tupBel1:0.186424):0.004886,(((((mm9:0.084505,rn4:0.091627):0.197835,dipOrd1:0.211666):0.022945,cavPor3:0.225634):0.010077,speTri1:0.148511):0.025643,(oryCun2:0.114421,ochPri2:0.201003):0.101624):0.015291):0.020683,(((vicPac1:0.107267,(turTru1:0.064676,bosTau4:0.123573):0.025145):0.040411,((equCab2:0.109311,(felCat3:0.098636,canFam2:0.102486):0.049838):0.006202,(myoLuc1:0.14262,pteVam1:0.113246):0.033792):0.004456):0.011576,(eriEur1:0.221758,sorAra1:0.269694):0.056557):0.021228):0.023628,(((loxAfr3:0.082165,proCap1:0.155353):0.026774,echTel1:0.246266):0.049887,(dasNov2:0.116609,choHof1:0.096318):0.053052):0.006229):0.399651,macEug1:0.133617):0.002474,monDom5:0.150921):

0.199105,ornAna1:0.461732):0.116917,((galGal3:0.164668,taeGut1:0.172833):0.200238,anoCar1:0.48763):0.10284):0.186338,xenTro2:0.834181):0.324842

Spanning 8.44 substitutions per site. We excluded 10-mers that after removing species with no alignable sequence due to either alignment gaps or missing sequence had a total branch length of less than 0.75. Data is available from

http://garberlab.umassmed.edu/data/conservation/hg19/omega/

For mouse we used the vertebrate multiple sequence alignment available from the UCSC genome browser for the mm10 assembly. We removed petMar1, gadMor1, oryLat2, gasAcu1, oreNil2, fr3, tetNig2, latCha1, xenTro3, chrPic1, anoCar2, melUnd1, taeGut1, melGal1, ornAna1, macEug2, sarHar1 vertebrate assemblies which left us with the following phylogeny:

((((((((((mm10:0.0861604,rn5:0.0923189):0.20235,dipOrd1:0.210872):0.0258938,(hetGla2:0.0916296,cavPor3:0.136929):0.0994423):0.00913482,speTri2:0.145406):0.0275377,(oryCun2:0.10975,ochPri2:0.200956):0.102105):0.0142197,(((((((((hg19:0.00672748,panTro4:0.00690586):0.00329132,gorGor3:0.00918574):0.00952813,ponAbe2:0.019182):0.00354391,nomLeu2:0.0218123):0.0117068,(rheMac3:0.00815625,papHam1:0.00799922):0.0289552):0.0208613,(calJac3:0.0342486,saiBol1:0.0333278):0.0358206):0.0593959,tarSyr1:0.137561):0.0111487,(micMur1:0.0919295,otoGar3:0.127188):0.0351183):

0.0153325,tupBel1:0.188903):0.0042042):0.0215023,((susScr3:0.121671,(vicPac1:0.10979,(turTru2:0.0635601,(oviAri1:0.0392014,bosTau7:0.0315737):0.0939007):0.0204197):0.00365643):0.0444426,((((felCat5:0.0897916,(canFam3:0.0888559,ailMel1:0.0767967):0.0218058):0.050101,equCab2:0.109329):0.00604713,(myoLuc2:0.137323,pteVam1:0.113957):0.0339856):0.00384687,(eriEur1:0.227177,sorAra1:0.270564):0.0629454):0.00322051):0.0291201):0.0231348,((((loxAfr3:0.0788116,proCap1:0.160315):0.00818092,echTel1:0.266806):0.00328658,triMan1:0.068537):0.0736006,(dasNov3:0.112113,choHof1:0.0974595):0.0536232):0.00734155):0.246266,monDom5:0.3541229999999999997):0.2125305,galGal4:0.5622546999999999):0.6482475,danRer7:0.871611):0.49907

Spanning 8.21 substitutions per site. We excluded 10-mers that after removing species with no alignable sequence due to either alignment gaps or missing sequence had a total branch length of less than 0.5. Data is available from

http://garberlab.umassmed.edu/data/conservation/mm10/mm10.omega

The models used were downloaded directly from UCSC and correspond to the alignments used.

II.16. Enhancer and Promoter Definition and Conservation Analysis

Enhancers and promoters were defined by H3K27ac peaks. We then merged all peaks from each time point located within 200bp from each other. Our maps consist of 28,142 and 29,273 H3K27ac regions (signal peaks) in mouse and human, respectively. We calculated the distance from each peak to the nearest transcription start site (TSS) of the highest expressed isoform for each gene using bedtools closest -D ref -t all (Quinlan and Hall, 2010). We classified all H3K27ac peaks that had a distance smaller than 500 bp to the nearest TSS as promoters, and the remaining peaks were considered enhancers. Enhancers were assigned to the nearest gene based on the same TSS distances as above. Unlike promoters, which were associated to the gene with the overlapping TSS independent of expression, enhancers were only associated to the closest expressed gene within 300 kb (Garber et al., 2012; González et al., 2015). This assignment of enhancers to nearby genes will misassign enhancers that either interact with more than one gene or interact with no adjacent genes. However, the majority of enhancers have been reported to interact with the neighboring gene (González et al., 2015). Overall, 2/3 of the peaks were annotated as enhancers in each species, consistent with previous studies (Villar et al., 2015). We filtered ATAC peaks to include only peaks that overlapped with a H3K27ac region. We classified ATAC peaks as enhancers or promoters based on the H3K27ac peak definition, and maintained the association to genes defined for H3K27ac peaks. To determine the conservation of mouse enhancer and

promoters in human, peaks were mapped to the human genome corresponding locations using liftOver -minMatch=0.1 -multiple (Hinrichs et al., 2006). We filtered out peaks that mapped to more than 3 locations and used the remaining peak locations to intersect with the human enhancer and promoter coordinates to determine if that region was also active in the human dendritic cells. To generate aggregation plots of the H3K27ac and ATAC-Seq signal, we used the center position of ATAC peaks for enhancers and the TSS for the genes associated to the peaks as coordinates for input to ngs.plot (Shen et al., 2014). The coverage was calculated for a 4kb region surrounding the center position (-L 2000). We selected the regions corresponding to each group of interest from the output matrix and calculated the mean signal per group.

II.17. ATAC and H3K27ac Dynamics

The mean signal across each ATAC-seq or H3K27ac peak was calculated by averaging the number of reads per base pair. The average signal across the libraries are normalized to the depth of each library using *DESeq2* (v1.10.1) in R (v3.3.1). ATAC-seq or H3K27ac peaks were considered dynamic in response to LPS if they have greater than two fold-change in their mean signal compared to unstimulated state. The dynamic ATAC-seq or H3K27ac peaks identified are clustered using k-means algorithm to identify groups of ATAC-seq H3K27ac peaks that are induced or repressed following LPS stimulation.

II.18. Motif Analysis

Motif analysis was done on 200 bp regions around the summits of the ATAC-seq peaks. The log-odds substitution rate for each 10 base-pair window across the summits of ECAs and ESPAs ATAC-seq peaks was calculated using SiPhy (Garber et al., 2009). The value of log-odd substitution score at the top ten percentile of a given peak was assigned as the conservation score for each peak. The kmers that intersected the ATAC-seq summits and which had log-odds score greater than 30 were considered for building cPWMs. To get a background set, we shuffled these 200bp ATAC-seq peaks within the enclosing H3K27ac peaks and considered all the kmers with log-odds score greater than 30. To identify kmers that distinguish the conserved ATAC peaks from background, we used the string kernel built-in gkm-svm R package (Ghandi et al., 2016) with 5 fold cross validation which resulted in 4500 unique kmers as features for conserved ATAC peaks. These kmers were clustered into 66 PWMs using k-medoids clustering algorithm with Euclidean distance, within the clara function in the *cluster* package in R (Blashfield, 1991). The cPWMs were then matched to the known motifs from CIS-BP database (Weirauch et al., 2014) using Tomtom (Gupta et al., 2007). Multiple motifs matched to the same TF are identified by numbers. For example JUN-1 and JUN-2. To find the cPWMs enriched in temporal gene groups or temporal ATAC peaks we used the Fisher exact test and all cPWMs with p value < 0.05 were considered enriched.

All cPWMs identified are available from

http://garberlab.umassmed.edu/publications/conserved_lexicon_Dec_2017/cPWMs.motifcPWMs.motif

II.19. Transposable Element Analysis

We used the transposable element annotation by RepeatMasker (Smit et al., 2004) to identify TE instances in each genome that overlapped at least 10% with the regulatory regions (enhancers and promoters) associated to induced genes. As a background, we shuffled these *cis*-regulatory regions in the genome inside boundaries defined by the regulatory regions associated to expressed genes with no response to LPS, expanded by 10kb in each direction. We then identified the number of instances for each TE family that overlapped at least 10% with these shuffled peaks. We performed this shuffling process 1000 times and compared the initial counts obtained for each TE family to this null distribution. We computed a p-value for this permutation and corrected it using the Benjamini Hochberg method. All TE families with adjusted p-value under 0.05 were considered to be overrepresented in the regulatory regions of induced genes. For each instance of these elements in induced genes, we identified the corresponding region in the other species' genome through *liftOver* as described above. We then evaluated if the region that can be identified in the other genome also overlaps a H3K27ac peak, classifying it as an ECA. H3K27ac and ATAC-Seq signal aggregation plots were generated as described above, with the TE start and end genomic coordinates as the target region, flanked by 1kb on each side.

II.20. Predictive Model of Gene Induction From cPWM Instances

Feature selection: For the selected set of 66 cPWMs, all instances were detected across all ATAC peaks (promoters and enhancers) using fimo (Grant et al., 2011), with a q-value threshold of 1e-4. We tested the models using two representations of the cPWMs as features: 1. All cPWM instances together - For each gene and each cPWM, we counted the number of instances across all regulatory elements of the gene. 2. All cPWM instances, separated by ATAC temporal pattern - each cPWM was separated to three features - the number of instances in LPS-induced regions (based on ATAC-seq data), number of instances in repressed regions and number of instances in unchanging regions.

Gene filtering: To build an informative model and to reduce noise from lowly expressed genes, we focused on highly expressed genes by taking only genes that were in the top 30% of expressed genes in at least one time point. Furthermore, to clearly distinguish induced from non-induced genes, we classified genes with a log2 fold change > 2 as induced, and genes with a log2 fold change between -0.3 and 0.3 as not induced, and discarded all the rest. Next, to create a balanced set of induced and non-induced genes, we downsampled the number of non-induced genes. This resulted in a total of 676 genes (338 induced and 338 non-induced) in mouse and 748 genes in human.

Model evaluation: All model training and evaluations was done in R, using the *caret* (v6.0.77) (Kuhn et al.) and *randomForest* (v4.6.12) (Liaw et al., 2002) packages. For each feature set, we evaluated the accuracy of the model

on the mouse data with 10-fold cross validation. For each one of the training data in the cross validation, hyperparameters tuning was performed using 10-fold inner cross validation with the “train” command, using the following parameters: tuneLength = 20, metric = “ROC”. To evaluate how well the model predicts induction on the human data, we trained a model on the full mouse data (again using 10-fold cross validation for hyperparameters selection) and applied the selected model on the human data.

Feature Importance: Importance measurement for each feature was computed with the “varImp” command, defined as the difference in mean accuracy across all trees between the model and the model after permuting the feature. The importance values were then scaled to span the range of 0 to 100.

II.21. Data and Software Availability

All samples generated for this work were submitted to NCBI as part of the Genomics of Gene Regulation Project, under accession number PRJNA356880. A list of samples used is specified on Table II-4.

II.22. Author Contributions

P.V. and E.D designed and performed the data analysis. S.A. and N.Y. designed and implemented the induced gene classifier. P.V. performed the mouse DC experiments and constructed the high-throughput sequencing libraries. B.T. developed the ATAC-Seq processing pipeline and advised in data analysis. S.M. performed human DC experiments. A.N. constructed the high-throughput sequencing libraries for human DCs. A.K. helped implement the data

processing pipelines and managed sample metadata. X.Z. designed and implemented the gene expression spectral clustering algorithm. W.D. and E.D. performed the TE analysis. P.M. supervised, developed protocols and planned all high-throughput sequencing experiments. M.G., J.L., and N.Y. conceived the project, advised on the analysis and data collection and supervised the research. P.V., E.D., and M.G. wrote the paper with input from all authors.

II.23. Acknowledgments

We want to thank Mitch Guttman, Jenny Chen, Ido Amit, Zhiping Weng, Scott Wolfe and members of the Garber Lab for valuable discussions and comments on the manuscript. We thank Idan Gabdank for help managing our data submission and to Sigrid Knemeyer for assistance with figures. This project was supported by the NHGRI U01 HG007910 (M.G., J.L., N.Y.), NIDA DP1DA034990 (J.L.), NIAID RO1AI111809 (M.G., J.L) and NCATS UL1 TR001453-02 (M.G.).

II.24. Declaration of Interests

The authors declare no competing interests.

II.25. Tables

II.26.Figures

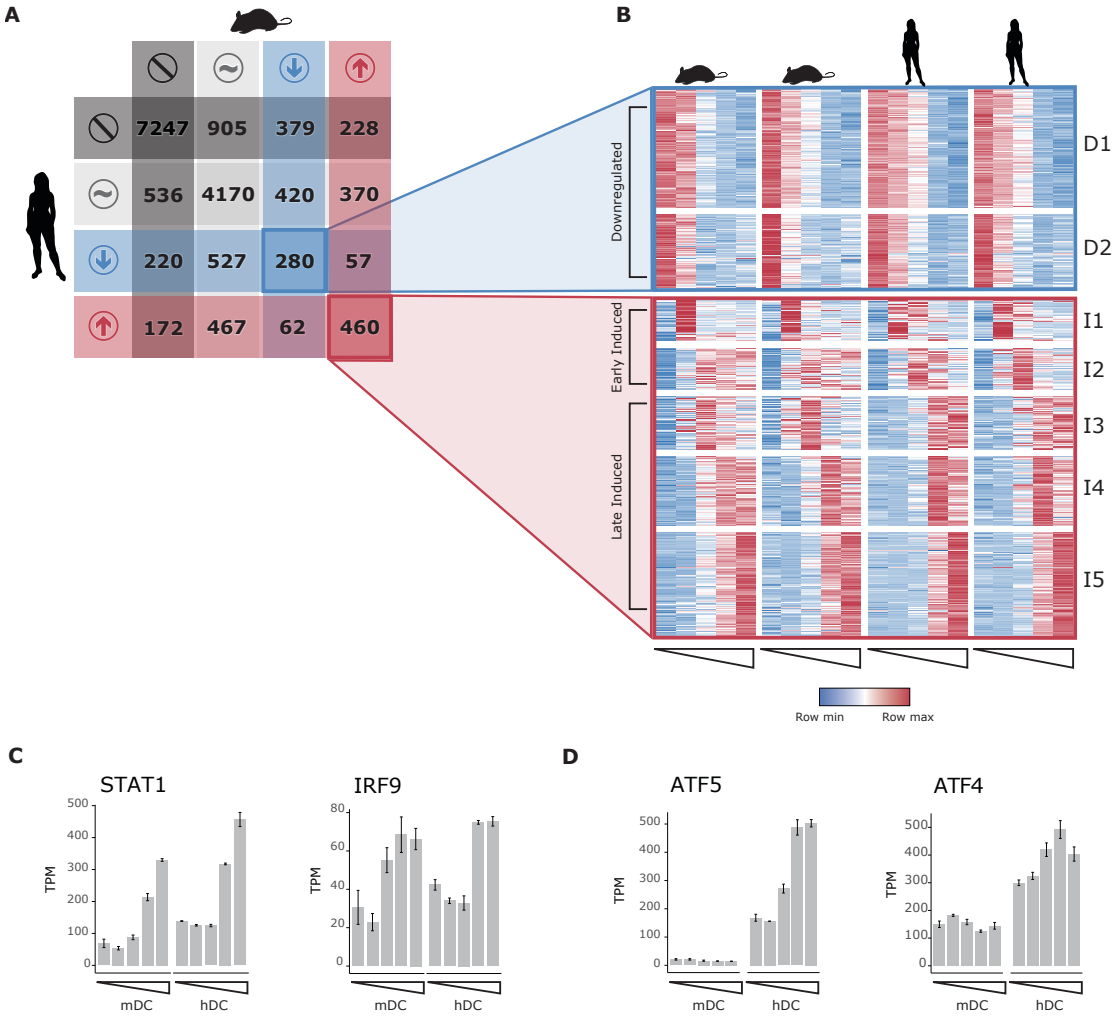


Figure II-1 | Highly induced LPS-responsive genes have similar expression kinetics in human and mouse dendritic cells

A) Classification of 16,500 homolog genes in mouse and human as not expressed (dark grey), expressed without significant change after LPS stimulation (light grey), downregulated (blue) or induced (red) **B)** Heatmap showing the normalized expression values for genes with shared response to LPS across five timepoints (Unstimulated, 1h, 2h, 4h and 6h post-LPS) in DCs derived from two different C57BL/6 mouse (left) and two human donors (right). Genes were grouped by spectral clustering into two clusters of shared downregulated genes (D1 and D2, top), and five clusters of shared induced genes (I1-I5, bottom). Induced gene clusters can be classified as early (clusters I1 and I2) or late (clusters I3, I4 and I5). **C)** Average normalized expression (TPM) in each time point for two examples of shared late induced transcription factors (TFs), Stat1 and Irf9. **D)** Average normalized expression (TPM) in each time point for ATF family TFs that respond differently in the two species.

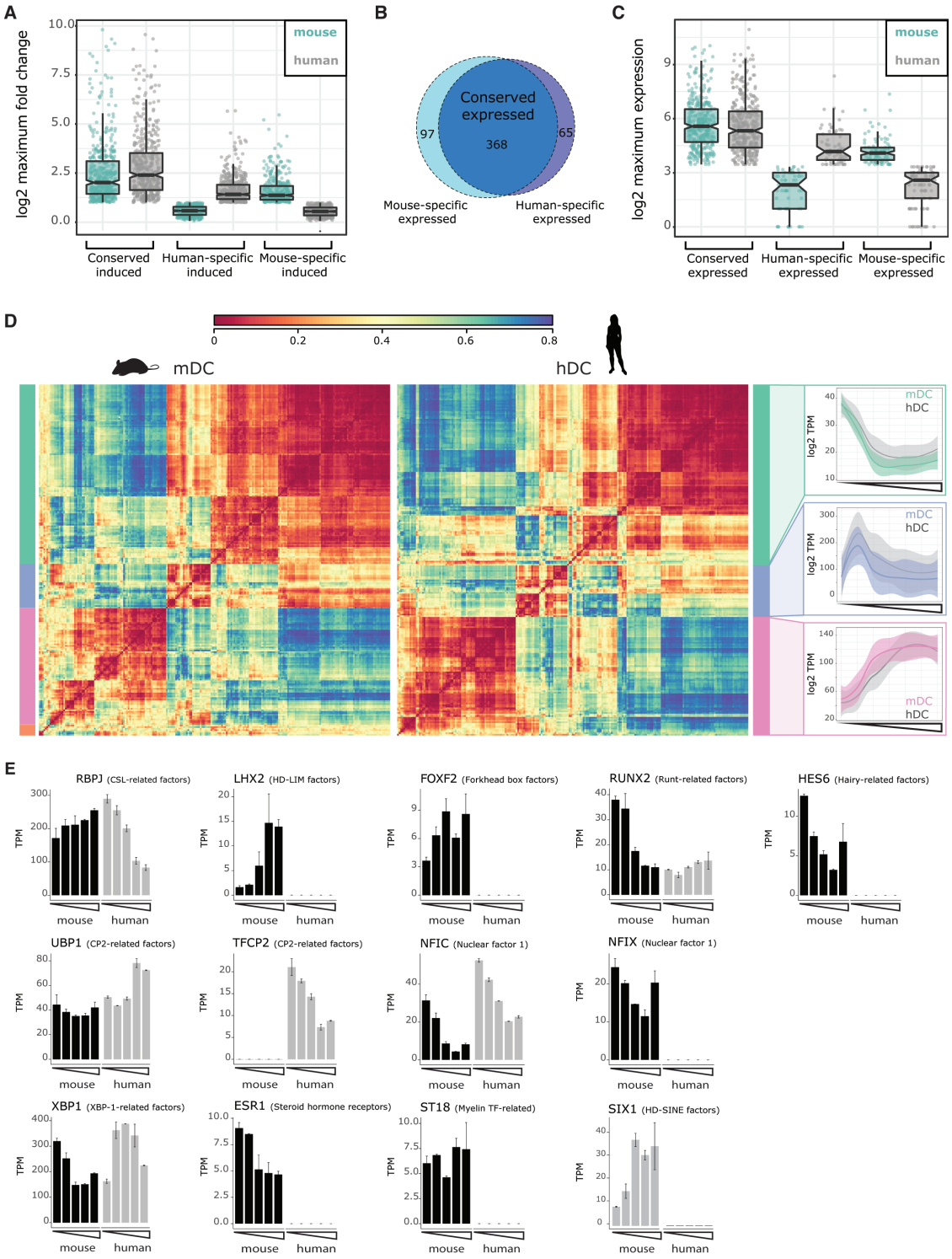

A
log2 maximum fold change
mouse
human
Conserved induced
Human-specific induced
Mouse-specific induced
B
Conserved expressed
97
368
65
Mouse-specific expressed
Human-specific expressed
C
log2 maximum expression
Conserved expressed
Human-specific expressed
Mouse-specific expressed
D
mDC
hDC
log2 TPM
E
RBPJ (CSL-related factors)
LHX2 (HD-LIM factors)
FOXF2 (Forkhead box factors)
RUNX2 (Runt-related factors)
HES6 (Hairy-related factors)
UBP1 (CP2-related factors)
TFCP2 (CP2-related factors)
NFIC (Nuclear factor 1)
NFIX (Nuclear factor 1)
XBP1 (XBP-1-related factors)
ESR1 (Steroid hormone receptors)
ST18 (Myelin TF-related)
SIX1 (HD-SINE factors)
TPM
mouse
human

Figure II-2 | TF network conservation in human and mouse DCs

A) Boxplots displaying log2 maximum fold change per gene for shared and species-specific induced genes post-LPS stimulation. **B)** Venn diagram of transcription factors expressed in each species. **C)** Boxplots displaying log2 maximum expression (TPM) for shared and species-specific expressed TFs **D)** Heatmap showing hierarchical clustering of the correlation of expression across time for all LPS-responsive transcription factors. Left: mouse factors (n=228); Center: human factors (n=224); Right: average expression of the factors in each cluster that show a shared pattern between species. **E)** Expression patterns of TFs that belong to families with only species-specific response to LPS.

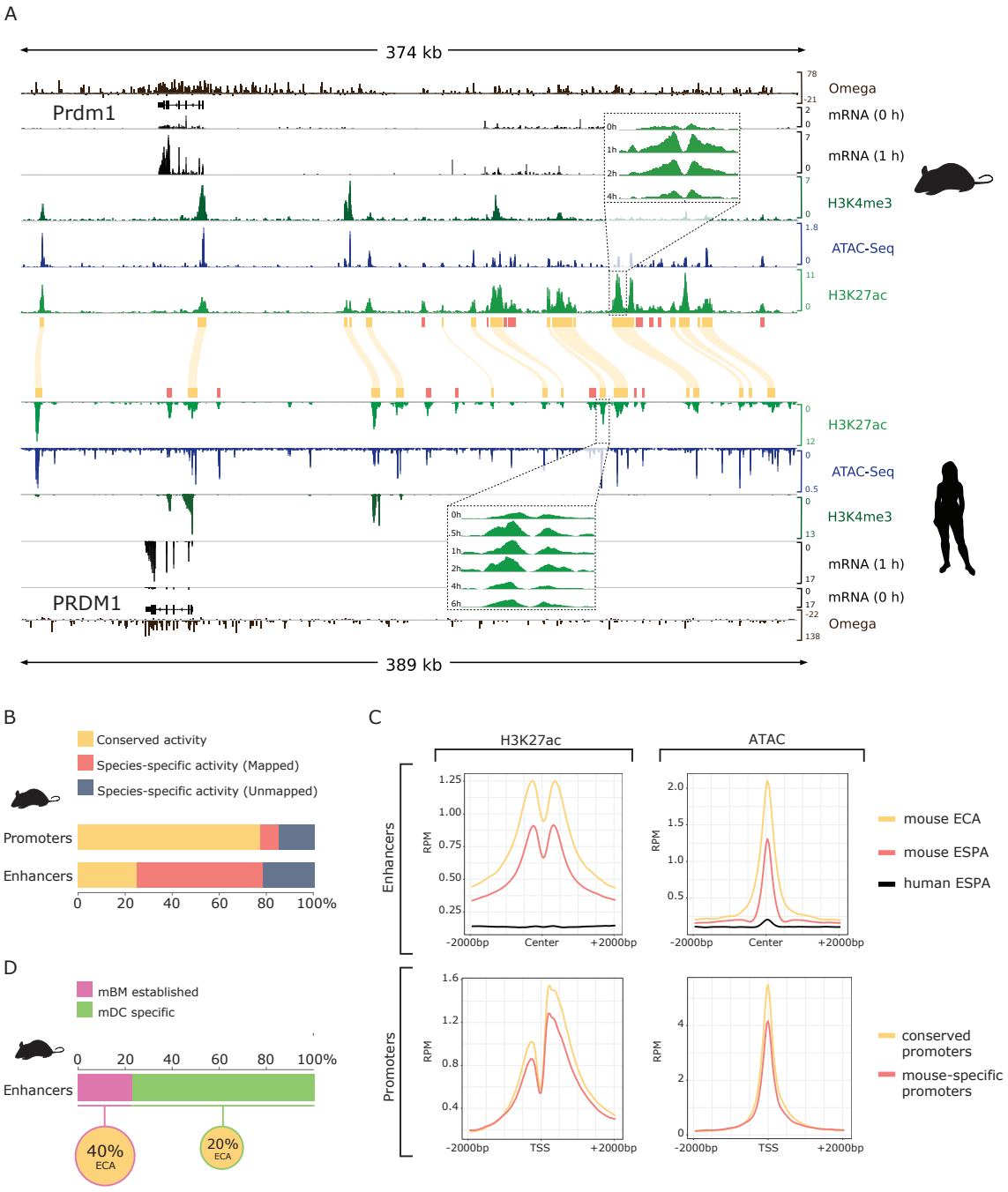


Figure II-3 | Rapid turnover of enhancer elements

A) Integrative Genome Viewer diagram of the PRDM1 regulatory region in both mouse (top) and human (bottom) displaying the data used in this study. Tracks

display from top to bottom: sequence conservation as estimated by SiPhy (Omega), RefSeq gene annotations, RNA-Seq coverage for unstimulated and 1 hour post-LPS, overlaid H3K4Me3, ATAC and H3K27ac coverage. The human data is in reverse orientation, yellow boxes and curved lines indicate conserved H3K27ac peaks (regulatory regions with conserved activity: promoters or ECAs). Inlets show the individual tracks for H3K27ac time course after LPS stimulation. Red boxes indicate H3K27ac peaks with species-specific activity. **B)** Proportion of regulatory regions with conserved activity: conserved promoters or ECAs, mouse-specific with clear human orthologous sequence (mapped promoters or ESPA) and mouse-specific with no clear orthologous sequence in human (unmapped promoters or ESPA) **C)** Average signal aggregation plots for mouse H3K27ac (left) and ATAC-Seq (right) signal over regulatory elements. Enhancer (top) H3K27ac signal is centered in open regions, defined by ATAC-Seq peaks. Promoter (bottom) H3K27ac is centered in the TSS. ATAC-Seq signal for enhancers is centered in open regions, while promoter ATAC-Seq signal is centered on the TSS. Data is shown for conserved enhancers and promoters (yellow), mouse-specific enhancers and promoters (red) and all other mouse genome coordinates for mapped human-specific enhancers and promoters (black). RPM = reads per million mapped reads **D)** Fraction of mouse enhancers that are already active (pre-established) in bone marrow (mBM) cells and enhancers that are mDC specific, and fraction of mBM pre-established or mDC specific enhancers that are conserved (ECA).

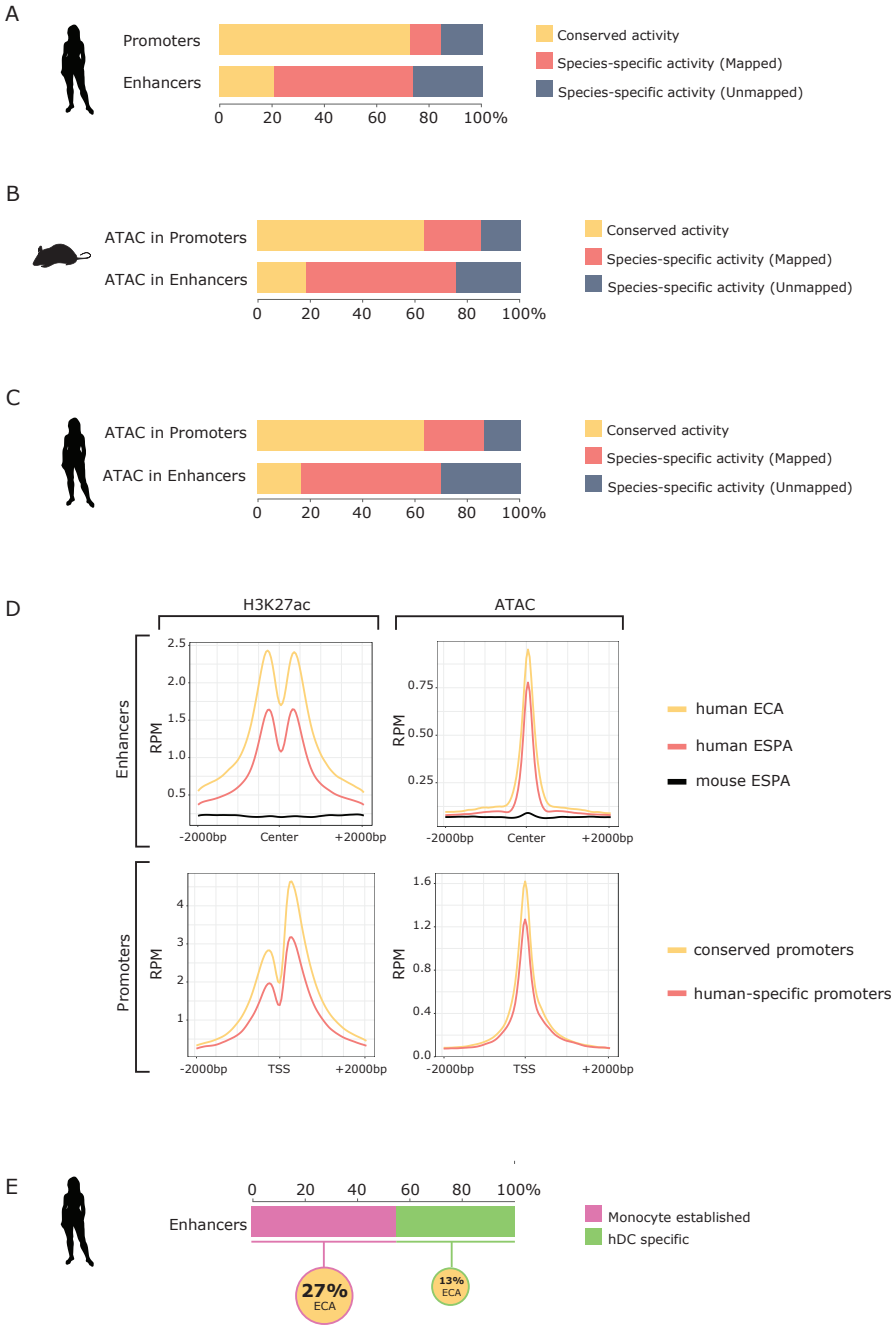
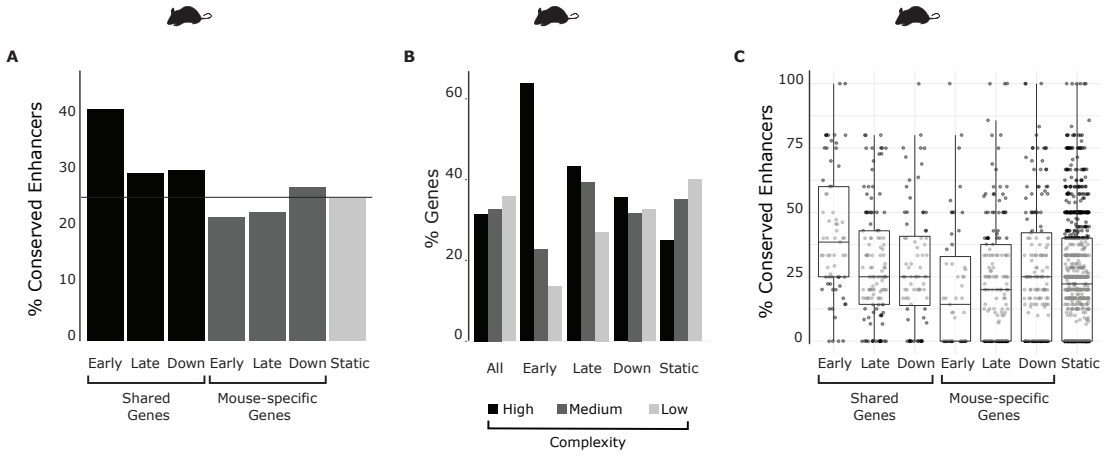


Figure II-4 | Conservation of enhancer promoter regions

A) Overall conservation of promoter and enhancer regions in human DCs. **B)** Overall conservation of ATAC peaks in promoter and enhancer regions in mouse DCs. **C)** Overall conservation of ATAC peaks in promoter and enhancer regions in human DCs. **D)** Average signal aggregation plots for human H3K27ac (left) and ATAC-Seq (right) signal over regulatory elements. Enhancer (top) H3K27ac signal is centered in open regions, defined by ATAC-Seq peaks. Promoter (bottom) H3K27ac is centered in the TSS. ATAC-Seq signal for both enhancers and promoters is centered in open regions. Data is shown for conserved regulatory regions (promoters or ECAs, yellow), human specific regulatory regions (promoters and ESPA, red) and all other human genome coordinates for mapped mouse-specific promoters or ESPAs. **E)** Fraction of human enhancers that are already active (pre-established) in monocytes (MONO) and enhancers that are hDC specific, and fraction of MONO pre-established or hDC specific enhancers that are conserved (ECA).



FigureII-5 | Genes with shared transcriptional response to LPS have complex regulatory loci and a higher conservation of enhancer activity in mouse.

A) Fraction of ECAs that are associated to genes that are early induced, late induced or downregulated upon stimulation with LPS in mouse DCs. The black horizontal line shows the average enhancer conservation for all genes B) Fraction of genes in temporal clusters that are associated to high-, medium- or low-complexity enhancer loci. C) Fraction of ECAs in high complexity genes that have shared or species-specific response. The response patterns are: early induced, late induced, downregulated or unchanged in response to LPS in mouse DCs.

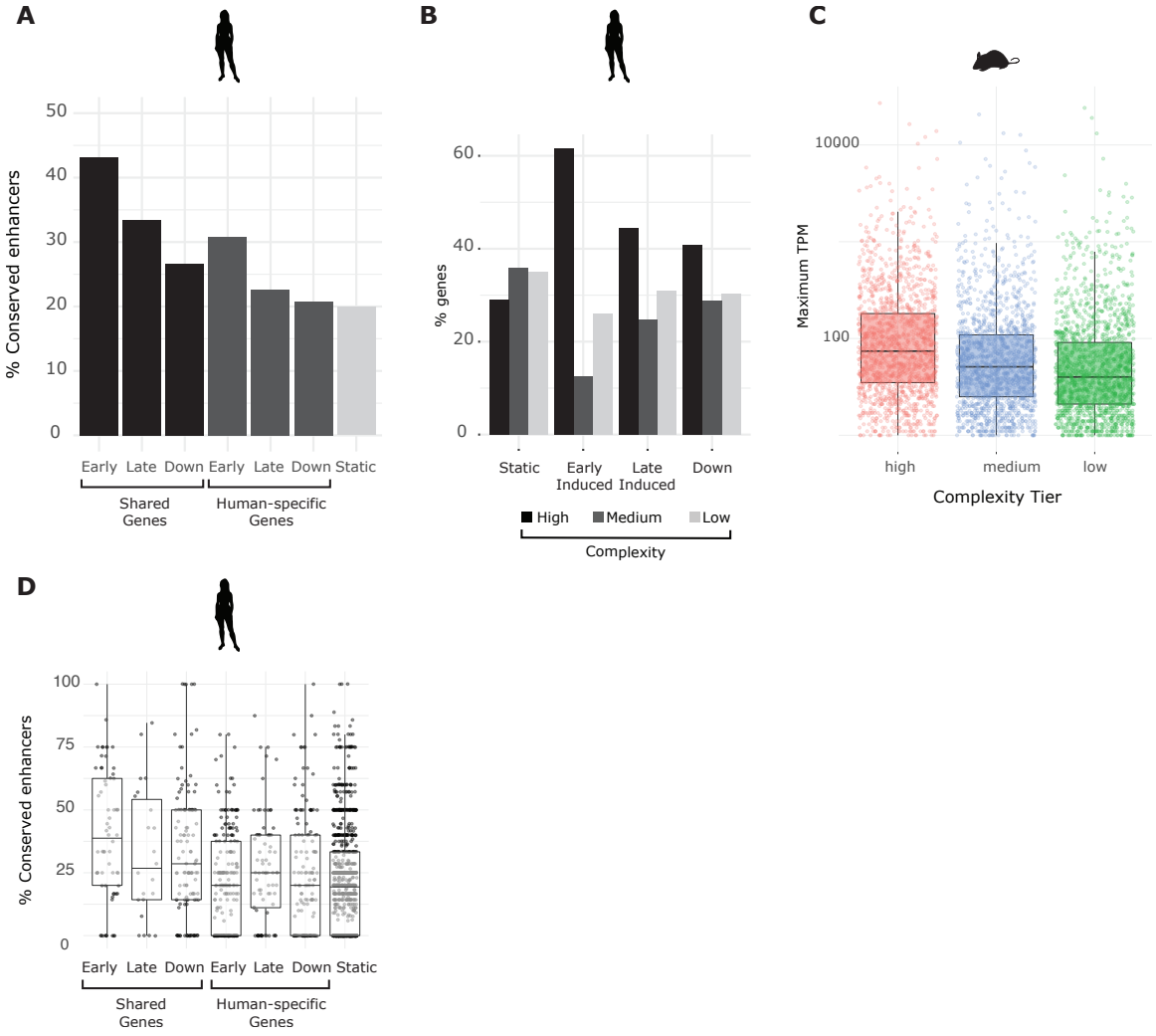


Figure II-6 | Genes with shared transcriptional response to LPS have complex regulatory loci and a higher conservation of enhancer activity in human.

A) Fraction of enhancers that are ECAs associated to genes that have shared or species-specific response: early induced, late induced or downregulated upon stimulation with LPS in human DCs. B) Fraction of genes in temporal gene clusters of human DCs that are associated to high-, medium- or low-complexity enhancer loci C) Maximum expression, measured in transcripts per million (TPM) for genes in each complexity tier D) Fraction of enhancers that are ECAs in high complexity shared or species-specific response genes which are early induced, late induced, downregulated or have no change in response to LPS in human DCs.

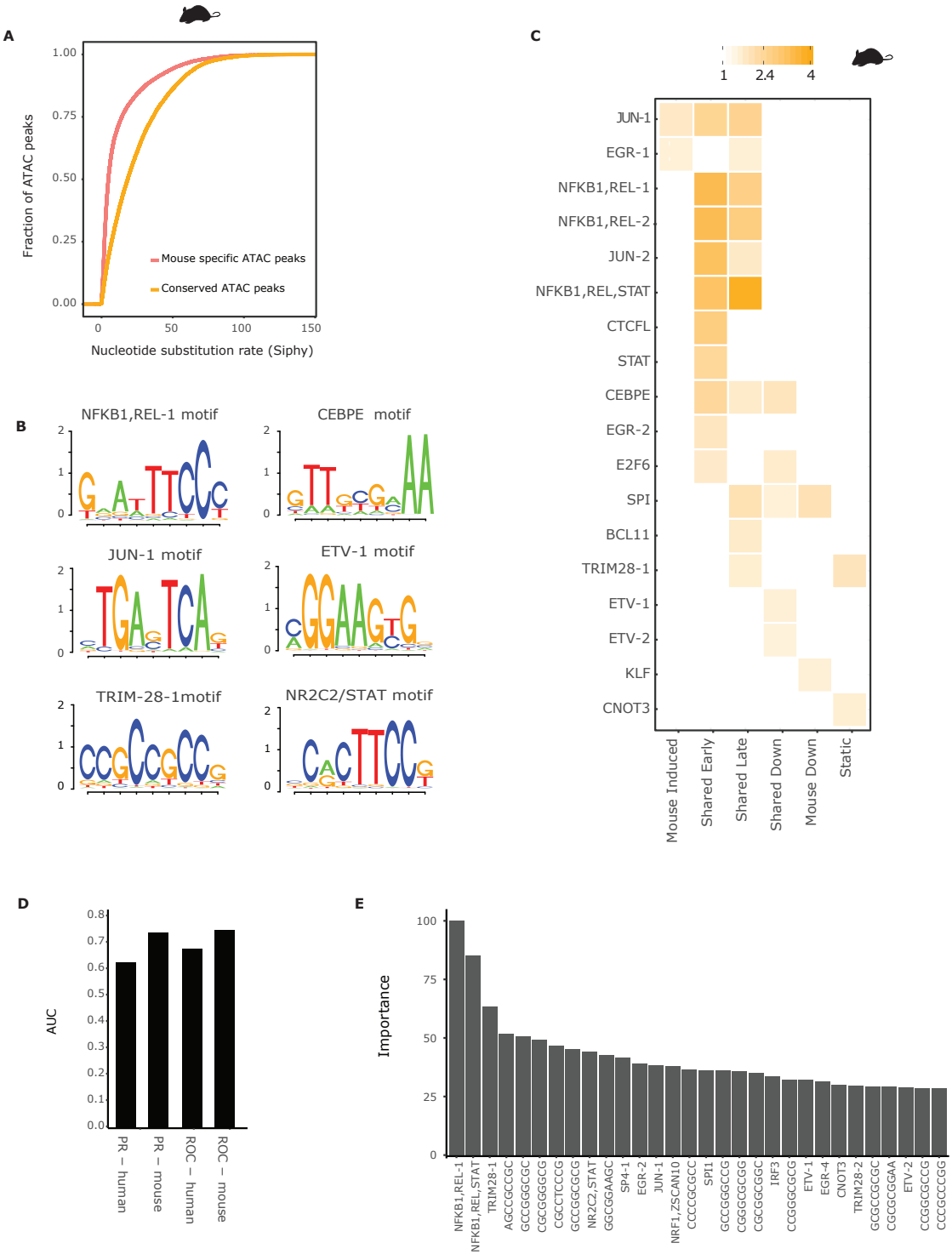

A
Fraction of ATAC peaks
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Nucleotide substitution rate (Siphy)
Mouse specific ATAC peaks
Conserved ATAC peaks
B
NFKB1,REL-1 motif
CEBPE motif
JUN-1 motif
ETV-1 motif
TRIM-28-1motif
NR2C2/STAT motif
C
1
2.4
4
JUN-1
EGR-1
NFKB1,REL-1
NFKB1,REL-2
JUN-2
NFKB1,REL,STAT
CTCFL
STAT
CEBPE
EGR-2
E2F6
SPI
BCL11
TRIM28-1
ETV-1
ETV-2
KLF
CNOT3
Mouse Induced
Shared Early
Shared Late
Shared Down
Mouse Down
Static
D
AUC
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PR – human
PR – mouse
ROC – human
ROC – mouse
E
Importance
100
75
50
25
0
NFKB1,REL-1
NFKB1,REL,STAT
TRIM28-1
AGCCGCCGCC
GCCGGGCGC
CGCGGGGCG
CGCCTCCCG
GCCGGCGCG
NR2C2,STAT
GGCGGAAGC
SP4-1
EGR-2
JUN-1
NRF1,ZSCAN10
CCCCGCGCC
SPI1
GCCGGGCCG
CGGGCGCGG
CGCGGCGGC
IRF3
CCGGGCGCG
ETV-1
EGR-4
CNOT3
TRIM28-2
GCGCCGCGC
CGCGCGGAA
ETV-2
CCGGCGCCG
CCCGCCCGG

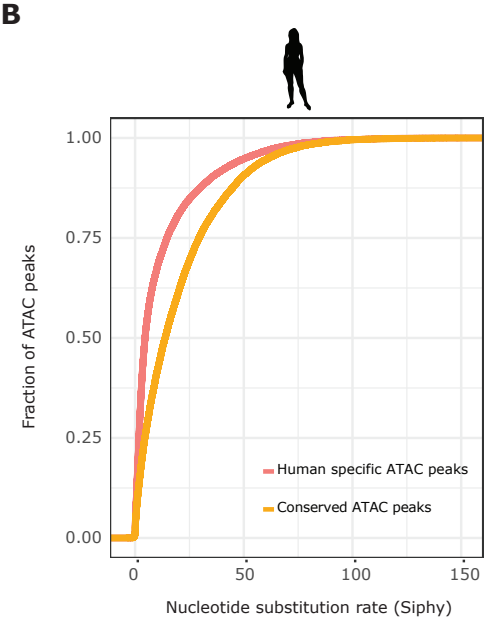
Figure II-7 | Enhancers with conserved activity contain a conserved lexicon

A) Distribution of SiPhy omega log-odds scores for 200bp regions around the summits of ATAC-seq peaks that have conserved signal (yellow) and species-specific signal (red) in mouse DCs. B) Examples of sequence logos of the clusters of kmers obtained after clustering the sequences in ATAC regions with conserved signal that have a log-odds score greater than 30. C) Enrichment heatmap showing the observed over expected values for each motif in ATAC-seq peaks with conserved signal associated to the gene groups defined in Figure II-1. D) AUC of the PR and ROC curves of a random forest model, predicting whether a gene will be induced or maintain constant expression following LPS stimulation. The features for each model were the number of each cPWM instances across all regulatory regions of each gene. E) Feature importance of the classifier. The importance of each feature was defined as the difference in mean accuracy across all trees between the model and the model after permuting the feature. The importance values were then scaled to span the range of 0 to 100. The 30 features with the highest importance values are presented.

A

	Number of peaks in ATAC	Total peaks in TF ChIP	Percent peaks in ATAC
Atf3	1497	1618	93%
Cebpb	23690	27348	87%
E2f1	2703	3690	73%
E2f4	925	1258	74%
Irf1	15375	17033	90%
Irf4	6222	6530	95%
JunB	7648	7843	98%
PU1	50576	64833	78%
Rela	19301	20966	92%
RelB	1475	1813	81%
Runx1	4098	4501	91%
Stat1	6085	6419	95%
Stat2	640	799	80%

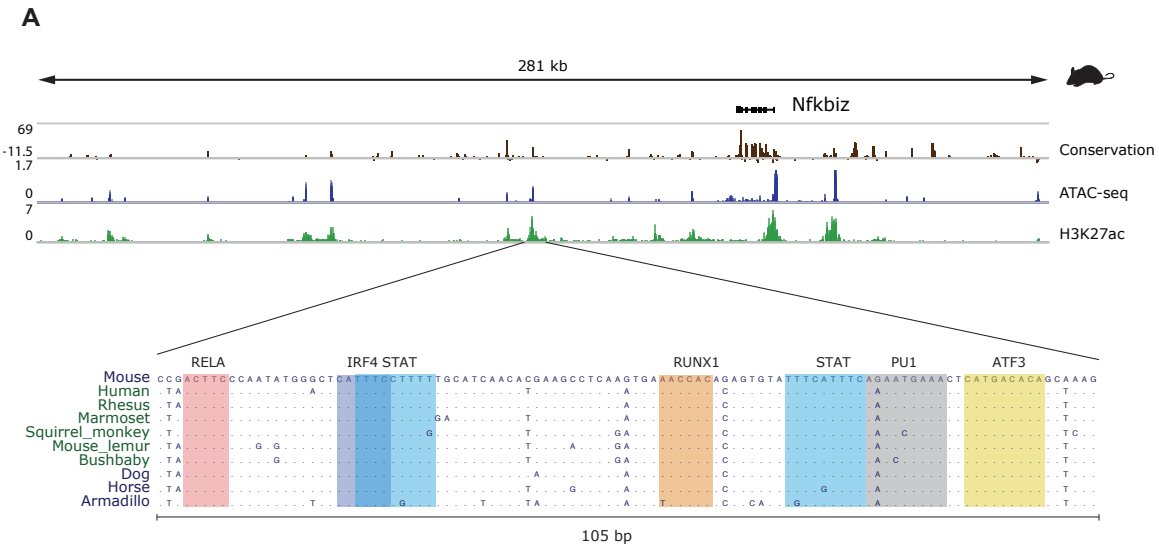
B



C

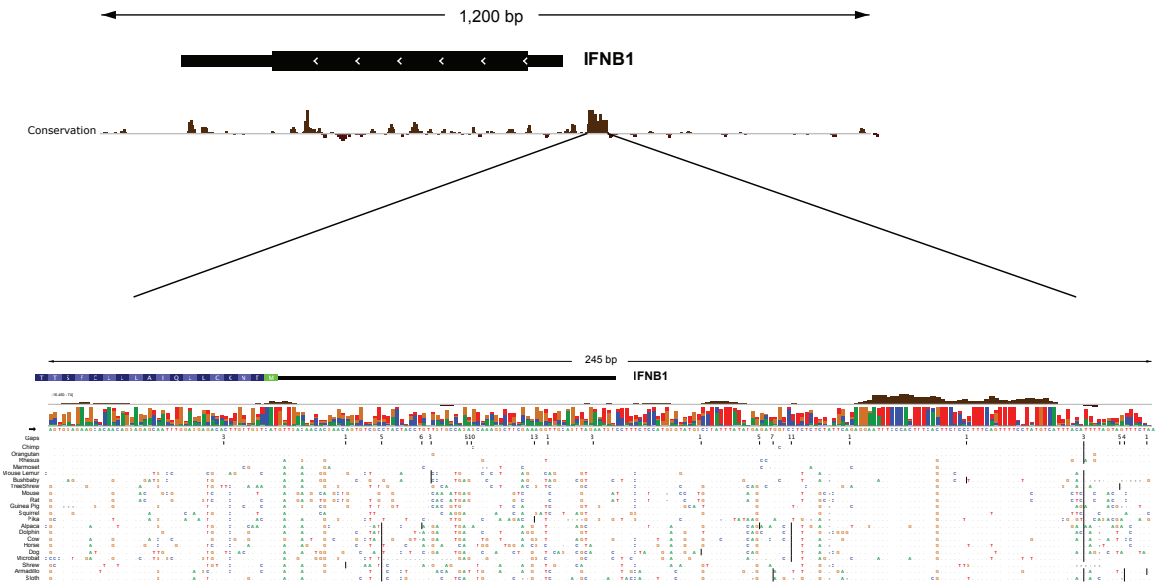
FigureII-8 | Sequence constraint of human ATAC- peaks

A) Table showing the number (column 1) and fraction (column 3) of TF ChIP peaks that are in ATAC-seq peaks B) Distribution of SiPhy omega log-odds scores in ATAC-seq peaks with conserved signal (yellow) and species-specific signal (red) for human DCs. C) Enrichment of cPWMs that are novel in the gene clusters.



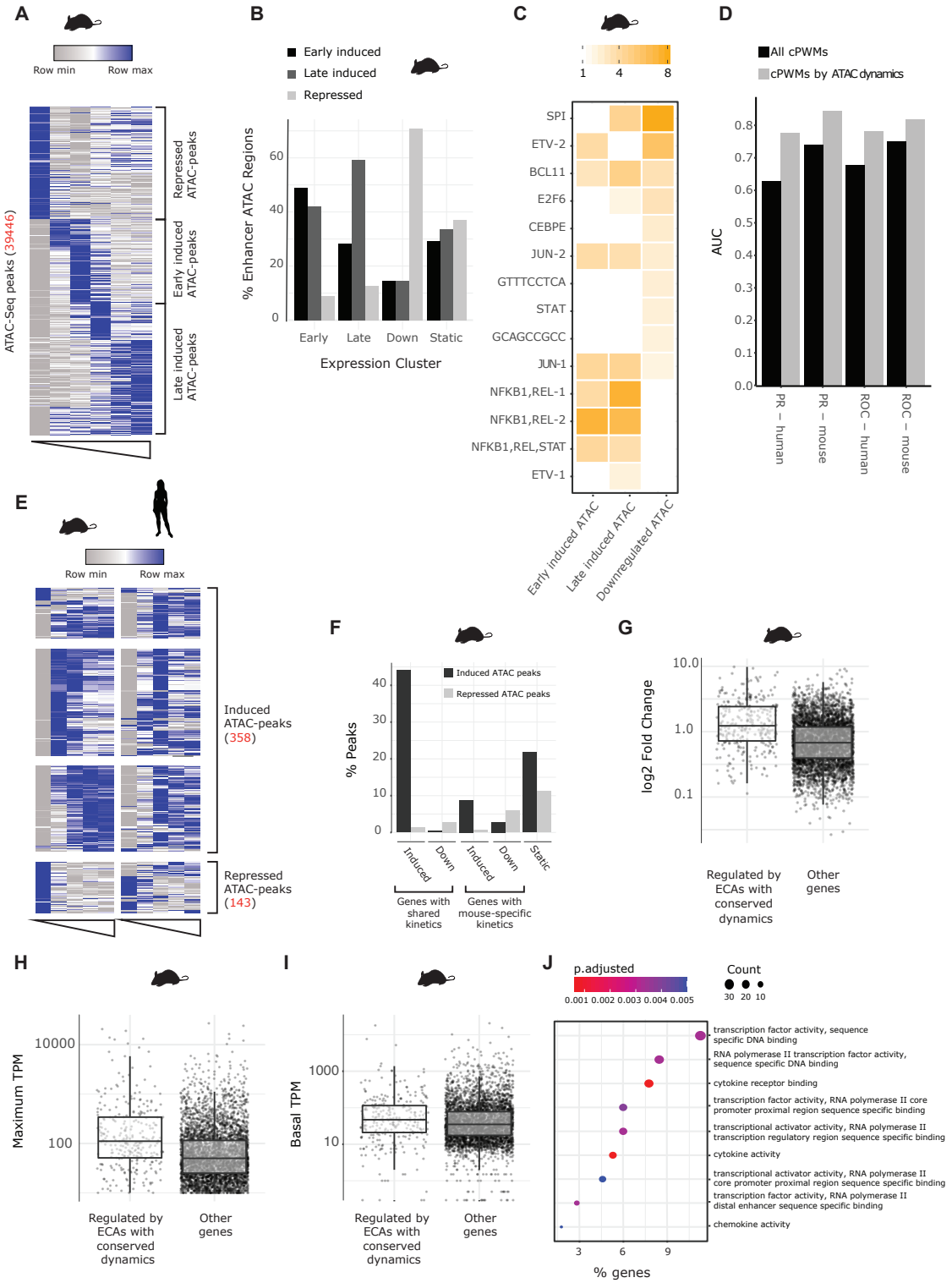
FigureII-9 | Candidate enhanceosome regions are highly conserved and bound by multiple TFs.

A) Example of an enhanceosome-like regulatory element in the NFKBIZ locus in mouse (top panel) showing the multiple sequence alignment of the conserved DNA accessible region



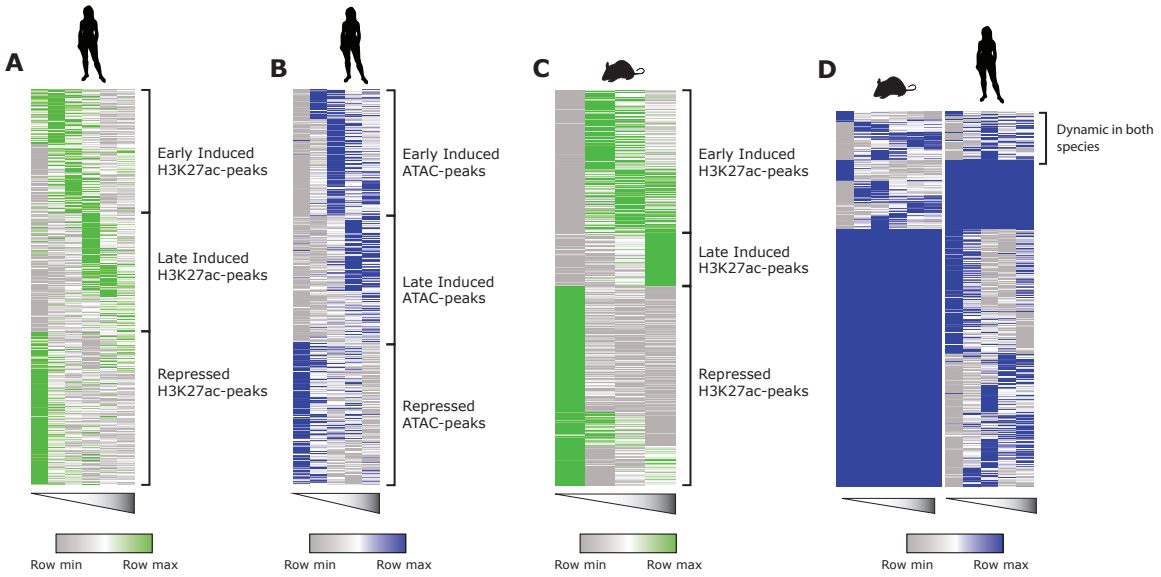
FigureII-10 | IFNb enhanceosome

IFNβ locus showing the multiple sequence alignment of the IFNβ enhanceosome and IFNβ gene. The top half of the figure shows the locus and conservation score, bottom half shows the multiple sequence alignment of the IFNβ enhanceosome. Dots are the nucleotides that haven't changed from the mouse sequence.


A
Row min
Row max
ATAC-Seq peaks (39446)
Repressed ATAC-peaks
Early induced ATAC-peaks
Late induced ATAC-peaks
B
Early induced
Late induced
Repressed
% Enhancer ATAC Regions
Early
Late
Down
Static
Expression Cluster
C
1
4
8
SPI
ETV-2
BCL11
E2F6
CEBPE
JUN-2
GTTTCCTCA
STAT
GCAGCCGCC
JUN-1
NFKB1,REL-1
NFKB1,REL-2
NFKB1,REL,STAT
ETV-1
Early induced ATAC
Late induced ATAC
Downregulated ATAC
D
All cPWMs
cPWMs by ATAC dynamics
AUC
PR – human
PR – mouse
ROC – human
ROC – mouse
E
Row min
Row max
Induced ATAC-peaks (358)
Repressed ATAC-peaks (143)
F
Induced ATAC peaks
Repressed ATAC peaks
% Peaks
Induced
Down
Induced
Down
Static
Genes with shared kinetics
Genes with mouse-specific kinetics
G
log2 Fold Change
Regulated by ECAs with conserved dynamics
Other genes
H
Maximum TPM
Regulated by ECAs with conserved dynamics
Other genes
I
Basal TPM
Regulated by ECAs with conserved dynamics
Other genes
J
p.adjusted
0.001 0.002 0.003 0.004 0.005
Count
30 20 10
transcription factor activity, sequence specific DNA binding
RNA polymerase II transcription factor activity, sequence specific DNA binding
cytokine receptor binding
transcription factor activity, RNA polymerase II core promoter proximal region sequence specific binding
transcriptional activator activity, RNA polymerase II transcription regulatory region sequence specific binding
cytokine activity
transcriptional activator activity, RNA polymerase II core promoter proximal region sequence specific binding
transcription factor activity, RNA polymerase II distal enhancer sequence specific binding
chemokine activity
% genes

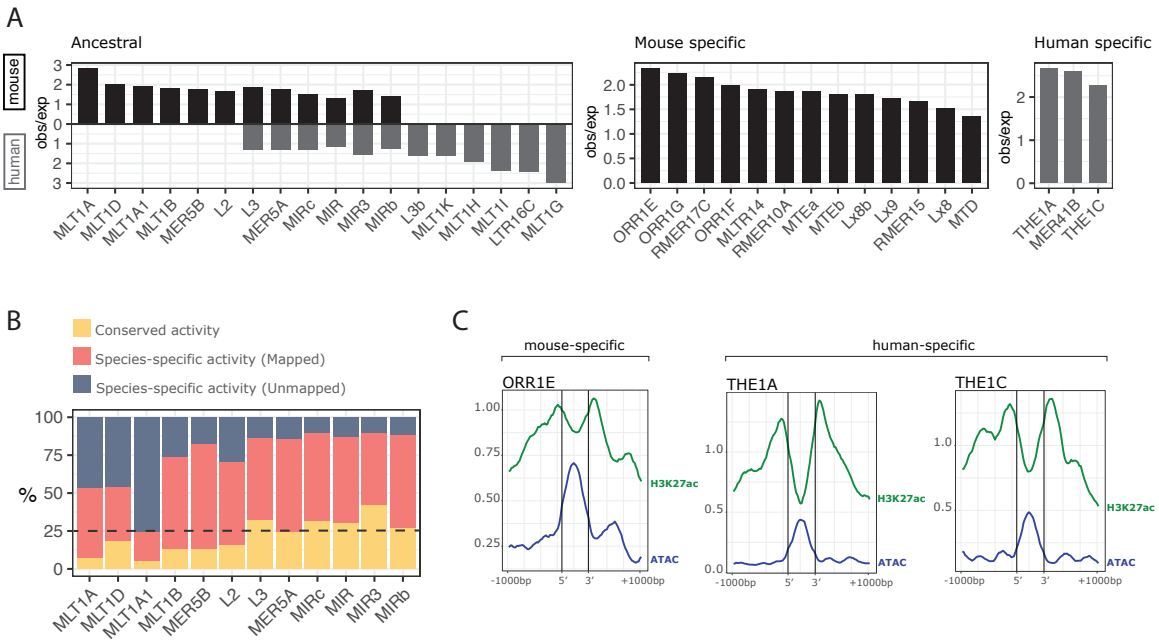
FigureII-11 | Regulatory regions with conserved activity and conserved kinetics regulate genes with shared induction kinetics.

A) Heatmap showing k-means clustered temporal patterns of mean signal per bp for ATAC-Seq peaks (at enhancer or promoter regions) with dynamic response to LPS in mouse DCs (Unstimulated, 30 minutes, 1 hours, 2 hours, 4 hours and 6 hours). Regions were classified as repressed, early induced or late induced. B) Fraction of early induced, late induced, downregulated or non-changing genes that are associated to dynamic ATAC peaks. C) Enrichment of cPWMs in ATAC peaks that are under purifying selection (Fig 4A) clustered into temporal groups. D) AUC of the PR and ROC curves of a random forest model, predicting whether a gene will be induced or maintain constant expression following LPS stimulation. The features for each model were the number of each cPWM instances across all regulatory regions of each gene (black bars), or all instances separated by the temporal pattern of the regulatory element (grey bars) E) Heatmap showing the temporal patterns of ATAC-seq peaks with conserved signal that are dynamic in both mouse and human. F) Enrichment of ATAC-seq peaks with conserved signal associated to genes that are induced in both mouse and human DCs, induced only in mouse DCs, downregulated in both mouse and human DCs, downregulated only in mouse DCs and not responsive to LPS in mouse DCs. G-I) The maximum absolute fold change, maximum tpm and baseline tpm of genes that are associated with ATAC-seq peaks with conserved signal that have same temporal response in both mouse and human versus all other genes J) Gene ontology analysis of genes associated with regulatory regions with conserved LPS response kinetics.



FigureII-12 | Dynamics of active regulatory elements and accessible regions

A) Heatmap showing the temporal patterns of H3K27ac peaks in response to LPS in human DCs (Unstimulated, 30 minutes, 1 hour, 2 hours, 4 hours and 6 hours) which are annotated as promoters or enhancers. B) Heatmap showing the temporal patterns of ATAC-seq signal associated with regions annotated as promoters or enhancers in human DCs (Unstimulated, 30 minutes, 2 hours, 4 hours and 6 hours) C) Heatmap showing the temporal patterns of H3K27ac peaks in response to LPS in mouse DCs (Unstimulated, 30 minutes, 1 hour and 2 hours) D) Temporal patterns of ATAC-seq peaks that are dynamic in at least one of the species when stimulated with LPS



FigureII-13 | Mobile elements of ancestral and recent origin have reshaped response to environmental stimulus.

A) Families of transposable elements (TEs) enriched in regulatory regions of induced genes in mouse and human. Observed over expected (obs/exp) values are shown for each TE only when the enrichment is significant in that species (p value < 0.004, permutation test; p adjusted < 0.05). Panels show families of TEs that have instances in the mouse and human genomes (Ancestral, Left), only in mouse (Mouse specific, Center), or only in human (Human specific, Right). B) Conservation rate of the enhancer regions that overlap each ancestral TE. C) Average signal aggregation plots of H3K27ac and ATAC-Seq over TE instances that overlap regulatory elements. Region is centered in each TE instance, delimited by the vertical bars, and the 2kb surrounding region is shown.

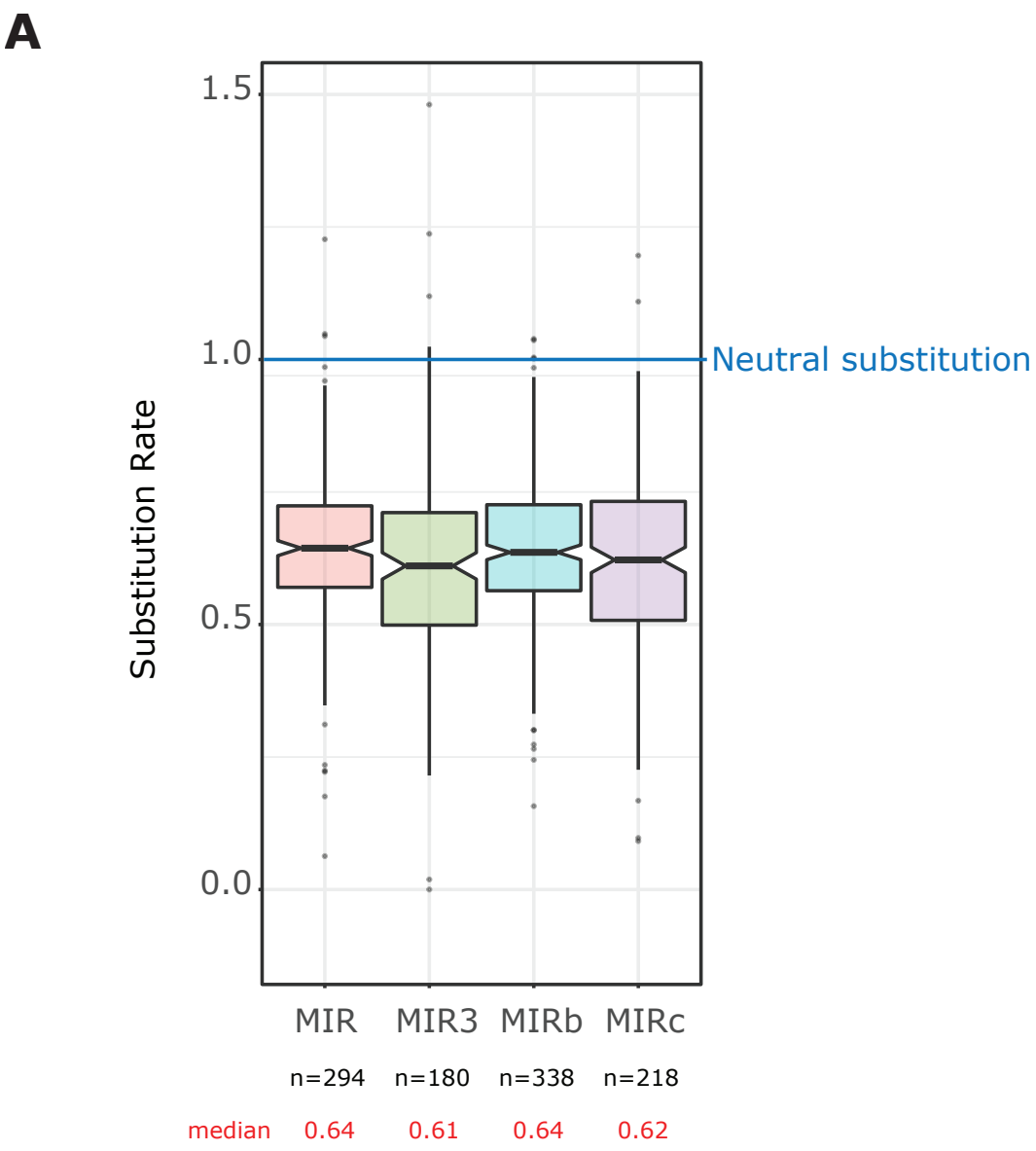


Figure II-14 | Nucleotide substituting rates of transposable elements

A) Distribution of the nucleotide substitution rates across 41 mammals for TEs from the MIR element families that overlap with regulatory regions of induced genes. One value for the substitution rate per element instance is shown, which corresponds to the value at the 90th percentile.

III. CHAPTER III: High-Resolution Mapping of Multi-Way Enhancer-Promoter Interactions Regulating Pathogen Detection

III.1. Preface

This research chapter encompassed work that is under revision in Molecular Cell, by Vangala Pranitha, Murphy Rachel, Quinodoz Sofia A., Gellatly, Kyle, McDonel, Patrick, Guttman, Mitchell, Garber Manuel. The publication is entitled “High-resolution mapping of multi-way enhancer-promoter interactions regulating pathogen detection”.

III.2. ABSTRACT

Gene regulation in eukaryotes involves thousands of distal elements. Understanding the contribution of enhancers to gene-expression is an unsolved problem that impedes our assessment of the role of risk variants within regulatory elements. We established a framework to tackle this problem by combining 3D enhancer-promoter (E-P) associations identified using Split Pool Recognition of Interactions by Tag Extension (SPRITE) with a predictive model of gene expression from DNA elements within associated enhancers. This framework dramatically outperformed models using E-P associations by genomic proximity alone and that these improved predictive models can be used to estimate the effect of enhancer loss in different divergent mouse strains. Further, we identified transcription factors that regulate the formation of chromatin interactions in

response to LPS. Finally using multi-way interactions inferred from SPRITE we found that genes that form stable enhancer-promoter hubs have less cell to cell variability in gene-expression as estimated from single-cell RNA-Seq data.

III.3. Introduction

Gene expression regulation involves a combination of promoters and distal regulatory elements, called enhancers (Dekker, Marti-Renom, and Mirny 2013; Schoenfelder and Fraser 2019; Snetkova and Skok 2018; Furlong and Levine 2018). Enhancers are thought to establish cell-type specific gene expression programs during development and in response to environmental cues. While there are on average six enhancer elements per promoter (Donnard et al. 2018; González, Setty, and Leslie 2015), it is still unclear how these enhancers coordinate gene expression.

Since the discovery that specific chromatin marks (H3K27ac and H3K4me1) were found to be enriched in enhancers regions (Tie et al. 2009; Creyghton et al. 2010; Calo and Wysocka 2013), tremendous effort has been made to annotate enhancers and associate them to their gene targets. These efforts resulted in a comprehensive catalog of putative enhancers elements across many cell types (Koch et al. 2007; Roadmap Epigenomics Consortium et al. 2015). Yet, we have not solved the fundamental question of how and to what extent enhancers contribute to the regulation of gene expression. For example, it is not currently possible to predict, given a set of active enhancers, what the spatiotemporal expression of a given gene may be.

There are several properties of enhancers that make predicting gene expression complex. First, while chromatin marks are correlated with enhancer activity it remains unclear whether all H3K27ac-rich regions represent functional enhancers (Pennacchio et al. 2013; Dickel et al. 2014). Second, enhancer-promoter (E-P) interactions can happen across tens of thousands of kilobases, sometimes skipping tens of intervening promoters (Ghavi-Helm et al. 2014; Claussnitzer et al. 2015). Third, many different may enhancers regulate a single gene, and each enhancer can regulate several genes (Donnard et al. 2018; González, Setty, and Leslie 2015; Ghavi-Helm et al. 2014) even simultaneously (Fukaya, Lim, and Levine 2016). Fourth, recent reports have shown that some E-P interactions may not bring regulatory elements into close proximity to their targets but instead can form phase-separated condensates of activators, co-activators and transcriptional machinery (Benabdallah et al. 2019).

Chromosome conformation capture methods (HiC, HiChIP, ChiAPet) have begun to uncover the complexity of eukaryotic gene regulation (Nott et al. 2019; Rubin et al. 2017; Song et al. 2019; Jeng et al. 2019; Mumbach et al. 2017, 2016; G. Li et al. 2012). However, these methods can only measure pairwise interactions across a cell population which complicates the discernment of multi-way interactions. It has also been reported that some interactions seem to occur across distances that are beyond the crosslinking distance of the current chromosome conformation methods and thus may be depleted in maps built

using these protocols (Benabdallah et al. 2019; Dekker 2016; Giorgetti and Heard 2016).

We recently described SPRITE (Quinodoz et al. 2018), a new framework for mapping higher-order spatial interactions in the nucleus. We showed that SPRITE can map long-range DNA contacts that cannot be observed by proximity ligation methods, and further SPRITE can map multiway contacts that occur simultaneously within a single cell. In this study, we sought to extend these observations to dissect regulatory contacts. We used SPRITE and established Sprite coupled with ImmunoPrecipitation (SIP) to define E-P interactions in Dendritic Cells (DCs) stimulated with Liposaccharide (LPS) with the goal of defining a predictive model of fold induction after this stimulation.

We found that interactions defined by SPRITE/SIP dramatically improved the performance of predictive models of gene induction even if less than half of H3K27ac regions interact with genes. Using multi-way analysis we find that E-P interactions involve multiple genes that form transcriptional hubs, and that DNA elements within these hubs synergize to enhance transcription. Interestingly the stability of these E-P hubs predicts the stability and consistency of expression across a cell population. These hubs can be dynamically organized during LPS response by key transcriptional regulators, such as Jun, Fos and other members of the AP1 family. Together, our results provide a comprehensive framework for understanding the quantitative contributions of E-P interactions to gene regulation, and how regulatory variants impact gene expression.

III.4. RESULTS:

III.4.I. SPRITE and SPRITE-IP Identify High-Resolution Chromosomal Interactions in Primary Bone Marrow-Derived Dendritic Cells.

Stimulating innate immune cells with TLR4 ligands results in highly reproducible temporal patterns of gene expression and epigenetic modifications. The temporal trajectories of genes and cis-regulatory elements (measured by H3K27ac ChIP signal and chromatin accessibility) follow four broad induction categories: early induction (1-2 hours post LPS stimulation), late induction (>2 hours post LPS stimulation), down-regulated, and non-changing (Amit et al. 2009; Garber et al. 2012; Bornstein et al. 2014; Donnard et al. 2018; Link et al. 2018). These clearly-defined modules of genes and their associated regulatory elements make this system well-suited to understand the nuances of the roles enhancers play in gene expression regulation (Amit et al. 2009; Garber et al. 2012; Bornstein et al. 2014; Donnard et al. 2018; Link et al. 2018).

To characterize regulatory interactions, we complemented previously generated expression and chromatin state maps (Donnard et al. 2018; Garber et al. 2012) with 3D chromatin interactions using SPRITE. We generated SPRITE interaction maps for BMDCs at 0hrs, 4hrs, and 24hrs following LPS stimulation (Figure III-1A). Each library was sequenced to ~400M reads, identifying ~50M unique clusters of interacting DNA molecules. To capture all the 3D interactions up to 24 hours following LPS stimulation, we merged the data from all 3-time points.

Manual inspection of interactions involving promoters of highly induced genes revealed that the unnormalized SPRITE signal was stronger on distal regions marked by H3K27ac which we refer to as active regulatory regions or enhancers (Figure III-1B,C). To test the generality of this observation we employed a viewpoint centric analysis. We centered our analysis on transcription start sites (TSSs) marked by H3K27ac, which we refer to as active promoters as defined recently (Donnard et al. 2018). For each active promoter we identified interactions that were overrepresented compared to a local neutral model that takes into account distance-based signal decay and genomic coverage (Methods).

This viewpoint analysis further strengthened our initial observations, confirming an enrichment of active regulatory regions among all promoter-interacting loci. This can be clearly appreciated at several distinct genes, where even distal regulatory elements display a strong promoter association. For example, an enhancer located within the Mtmr7 gene shows spatial proximity only to the promoter of the Slc7a2 gene even though there are 5 intervening genes between Mtmr7 and Slc7a2. Notably, no other expressed gene in this locus interacts with this enhancer (Figure III-1D). Globally, when taking active promoters as viewpoints, the SPRITE signal was enriched for active regulatory regions compared to non-active regions (Figure III-2A). Using this signal we were able to connect 6,636 (76%) active promoters with ~9,000 (40%) enhancers in these cells. (Figure III-1E left).

For the majority (60%) of enhancers we did not detect interactions with an active promoter, this could be due to several possibilities: 1) these putative enhancers do not regulate gene expression in steady-state DCs or in response to TLR4 signaling, 2) lack of connectivity could result from lack of power to detect interactions using the SPRITE assay at this depth or 3) Enhancers may be acting from a distance that is beyond what SPRITE can capture (Benabdallah et al. 2019). To help identify which of these phenomena explains the overall lack of interactions, we set out to determine the classification ability of these enhancers (Methods). We built random forests using transcription factor binding motifs (TFBMs) to predict if a gene is induced or not using three different sets of enhancers (Figure III-1E middle): 1) enhancers that were 300 Kb from a promoter 2) enhancers associated by SPRITE and 3) enhancers that are assigned to promoters based on a linear proximity of <=300kb, but that have no enriched signal in SPRITE. We found that the area under the curve (AUC) of a classifier built with TFBMs within the enhancers for which there is evidence of physical proximity by SPRITE to a promoter improved to 0.79 compared to 0.69 for the classifier based on linear proximity (<=300kb on both sides) (Figure III-1E right, III-2B). Moreover, the classifier trained using TFBMs within enhancers that are not connected by SPRITE has almost no predictive power. Notably, the most influential TFBMs for the most predictive classifier are the DNA binding motifs of transcription factors (TFs) that are well-characterized key regulators of LPS-induced genes (Medzhitov and Horng 2009) (Figure III-2C). This suggests that

the model built on our enhancer assignments captured the biologically relevant signal and that unconnected enhancers introduce noise that impacts the performance of a classifier.

To increase the sensitivity to detect E-P interactions, we coupled SPRITE with immunoprecipitation in a method we called SPRITE-IP (SIP) (Figure III-3A). As expected, the resulting libraries are highly enriched for H3K4me3-marked transcription start sites (TSSs), an enrichment that is absent in SPRITE data (Figure III-4A) and SIP libraries are three-fold enriched for clusters containing promoters. We compared SIP to HiChIP, a similar enrichment protocol, by performing SIP for H3K4me3 in mESC, for which there is abundant HiChIP data. We found that 72% of SIP interactions are also identified in previously published H3K27ac-HiChIP (Figure III-4B). Our analysis recapitulated most (4 of the 5) of previously identified interactions from 3C/4C methods (Table III-1 sheet1) and 9 out 11 (Table III-1 sheet2) experimentally validated distal enhancer interactions including the enhancer associated with Sox2 promoter (Figure III-4C) (Y. Li et al. 2014; Moorthy et al. 2017; Juric et al. 2019). Manual inspection of several loci showed that similar interactions are enriched in both SIP and SPRITE libraries (Figure III-3B). Overall, we find that SIP not only recovered the vast majority (~74%) of E-P interactions identified by SPRITE (Figure III-3C) at higher frequency but also yielded a significant increase in new interactions. We concluded that SIP enriched signal involving promoter interactions without introducing any detectable bias.

We merged our SIP and SPRITE data, which together represent 1 billion unique reads with 78M medium-size clusters (2-100 unique DNA molecules) across all three-time points (0hrs, 4hrs, 24hrs). With this data, we detected 34,187 interactions involving 7,130 active promoters (86% of all active promoters in BMDCs) (Table III-2). The 1,122 genes with active promoters for which still could not detect interactions had low expression or were not highly induced by LPS compared to genes for which we detected interactions (Figure III-4D).

III.5. Dramatic Improvement in Predicting Gene Expression

One way to estimate the accuracy of associating enhancers to promoters is by evaluating the predictive power of a model that estimates the change in gene expression from the TFBMs within the associated enhancers. Surprisingly, associating enhancers to their neighboring genes resulted in models that poorly predict expression (Figure III-5A,B); however, previous studies showed that iterative reassignment of enhancers of inactive genes to active genes can dramatically improve model performance (González, Setty, and Leslie 2015). This suggests that gene expression can be predicted from the regulatory sequence, but proper enhancer association is required.

We next sought to predict gene expression changes from TFBMs using different approaches to associate TFBMs to genes. We assigned a score to each TFBM within accessible regions of enhancers and promoters using SeqGL (Setty and Leslie 2015), a score that reflects the importance of the TFBM to distinguish the enhancer from non-enhancer region and used the sum of these scores

across all regulatory elements with similar DNA accessibility kinetics that are associated with a gene as features in an elastic net regression model to predict changes in gene expression after LPS stimulation (Methods).

We first built regression models where TFBM scores were combined for enhancers 300Kb of a gene promoter. The model predicting gene induction after one hour of LPS stimulation had only a 0.3 Spearman rank correlation () while the model predicting six-hour induction had a =0.21(Figure III-5A,B). The model performance did not improve with different neighborhood windows (Supplement Note). We then built models using only TFBMs within enhancers for which there was 3D interaction evidence. We observed a dramatic improvement in model performance when we used 3D-defined interactions. Indeed, predicting the induction levels at one and six hours after LPS stimulation dramatically improved with = 0.72 and =0.65respectively (Figure III-5C,D) (Methods). Taken together, these results show that assigning enhancers based on interaction data, notably improves both the specificity and sensitivity of the predictive models of gene expression and emphasizes the importance of accurate regulatory element assignment to predict gene expression from TFBMs alone.

Upon examining the features with non-zero coefficients, we found that transcription factors such as REL, STAT, SPI, CEBP have positive coefficients supporting their known role as activators in the first hour following LPS stimulation, but interestingly negative coefficients for PRDM1, RUNX1 transcription factor motifs suggest a repressive function for these factors (Figure

III-6A). Similarly, the 6-hour model points to STAT and SPI as the strongest activators while also highlighting PRDM1 and RUNX family members as transcriptional repressors (Figure III-6B).

It is important to note that removing the dependence on the timing of DNA accessibility reduced the predictive power of the model (Figure III-6C), while scoring schemes that used the intensity of either the ATAC or H3K27ac signal did not improve the model performance (Figure III-6D, Supplemental Note).

III.6. Induced Genes Form Transcriptional Hubs

A large fraction of interactions in our dataset (35%, 11,953) represents direct contacts between two gene promoters (P-P interactions). In the vast majority (95%) of P-P interactions one of the promoters belongs to an expressed gene (>10 TPMs), and in the majority (66%, 7,530) both promoters belong to expressed genes (Figure III-5E).

It has been observed that genes with physically interacting promoters tend to have similar expression levels in K562 cells, indicating that they were co-regulated (G. Li et al. 2012). Accordingly, we observed that LPS-induced genes with interacting promoters tend to have higher temporal correlation than gene pairs with no interacting promoters; this is true even after controlling for expression level, the number of enhancers interacting with individual promoters, and genomic context (p-value < 10^{-15}, Wilcox rank-sum test, Figure III-5F, Methods). Consistent with this observed co-regulation our model significantly worsens when TFBMs within interacting promoters are excluded (=0.51versus

=0.65, Figure III-5G) when including them, Figure III-5B). This further strengthens the importance of transcriptional hubs that result in synergistic interactions between co-regulated genes (Furlong and Levine 2018; Rieder, Trajanoski, and McNally 2012; Edelman and Fraser 2012).

Some gene hubs include critical cytokines that are highly induced by LPS. For example, the chemoattractants Cxcl1-3 on chromosome 5 have highly correlated induction (Figure III-5H). Manual inspection of the enhancer connections in this region revealed that many enhancers established connections with more than one of these three chemokine genes simultaneously in the same cell but not with the other genes in the region that are not induced (Figure III-5I). To further explore the mechanism that results in this coordinated expression, we used the ability of SPRITE to capture complex genomic interactions involving multiple DNA elements at a single-molecule resolution to investigate the architecture of transcriptional hubs. There are on average 6!=720 possible configurations per promoter. Given this large number possible configurations, we reasoned that a clustering approach would reveal complex interactions whenever they are frequent in the cell population. We employed a biclustering approach, in which we co-clustered genomic loci and SPRITE molecules simultaneously to identify non-adjacent regions involving at least one promoter (Figure III-5J,K). We then determined which interactions occurred at a frequency higher than expected under a null model where we assumed that regions interact with equal frequency (Methods). At a false discovery rate (FDR) of 5%, we found that 1895 P-P pairs

that included two expressed promoters and at least one enhancer (1619) or a third promoter (276). When two expressed promoters form a hub with a shared enhancer, they tend to be more correlated than when the interactions involved only two promoters and no enhancer (Figure III-5I). This supports the notion of enhancer hubs as means to coordinate gene expression and is consistent with a recent study that showed that placing an enhancer between two active promoters led to coordinated gene expression (Fukaya, Lim, and Levine 2016).

III.7. Variability in Regulatory Configurations Correlates With Stochasticity in Gene Expression:

Given that genes within stable hubs tended to be correlated, we also reasoned that such genes may also have a more stable expression. We generated single-cell RNA-seq (scRNA-Seq) libraries for BMDCs at 0hrs, 1hrs, and 4hrs post LPS stimulation. Manual inspection showed that genes in different categories: Induced, unresponsive, or downregulated had different noise levels even after accounting for expression levels. For example, Rps28 and Gnai2 are not responsive to LPS and are expressed at very similar levels, however, Rps28 has a lower variation at transcript level than the Gnai2 (Figure III-7A). The inspection of bi-clusters for both genes showed that Rps28 had clear bi-clusters while none are detectable for Gnai2 (Figure III-7B,C). To test the hypothesis that stable hubs reduce cell to cell expression heterogeneity, we focused on highly expressed genes (bulk RNAseq data max TPM > 100) for which we could reliably estimate variability from scRNA-Seq. We also required that genes were

associated with six or more enhancers, representing a highly complex regulatory landscape (Donnard et al. 2018; González, Setty, and Leslie 2015). After this filtering, our analysis concentrated on 2038 genes, of which we could identify bi-clusters for 638 while the remaining 1400 genes had only detectable pairwise interactions. Regardless of LPS responsiveness, we found that genes for which we find bi-clusters have lower variability in expression as estimated from scRNA-Seq when compared to genes without higher-order configurations (Figure III-7D,E).

III.8. Quantitative Induction Predictive Model Identifies the Effect of Enhancer Loss

A recent study explored the impact of naturally occurring variation on the transcriptional response in bone marrow-derived macrophages (BMDMs) stimulated with Kdo(2)-lipid A (KLA), a TLR4 agonist analogous to LPS, in 5 different strains of inbred mice (Link et al. 2018). Because promoter changes could not explain the large differences between LPS inducible genes between strains, the authors hypothesized that distal regulatory elements, rather than promoters, are responsible for these changes. We sought to test this hypothesis in the context of our 3D interaction data. We note that our data was generated in BMDCs, while Link et al. used BMDMs. However, our comparison of the transcriptional response to LPS showed that both cell types have a remarkably similar transcriptional response to TLR4 ligands (Figure III-8A) and both cell types share a very high fraction (70%) of H3K27ac-rich regions (Figure III-8B).

We focused on BMDMs of C57BL/6J and SPRET/EiJ strains. These two strains are highly divergent with an average of one single nucleotide polymorphism (SNP) every 120bp. BMDMs from the two strains have 4,820 (p-adjusted <0.01) differentially expressed genes and 11,345 H3K27ac-rich regions that significantly differ in their activity (measured by H3K27ac ChIP-Seq signal, Methods) at baseline or post KLA stimulation. To test our model we focused on genes that were induced at different levels between the two strains.

For example, Cxcl10 (a critical T cell chemoattractant) is induced 2.8 times more in C57BL/6 than in SPRET/EiJ BMDMs (Figure III-7F). All H3K27ac-rich enhancers within this locus are shared between C57BL/6 and SPRET/EiJ except for one enhancer that is lost in SPRET/EiJ BMDMs (Figure III-7G, black box). Our interaction data shows that although the enhancer is 100 Kb away, it physically interacts with Cxcl10 in C57BL/6. Interestingly the same enhancer also interacts with the adjacent gene N-Acylethanolamine Acid Amidase (Naaa), and Naaa also shows lower induction in SPRET/EiJ (Figure III-7F). This strongly suggests that this enhancer may be important to the higher induction levels observed in C57BL/6. To extend this observation we next focused on enhancer regions that showed differential H3K27ac signal between the strains and for which we could identify significant interactions with at least one promoter (5,449 of the 11,345 differential enhancers). Remarkably, genome-wide majority of the interactions of enhancers with differential H3K27ac signals are with genes that are differentially expressed between strains, compared to genes that are not

differentially expressed (62% vs 38%, fisher's p-value < 10^{-6}). The observed enrichment is not the result of any genomic or spatial configuration as we do not see any such enrichment when we randomly assign enhancers to neighboring promoters (Methods).

To better understand the importance of enhancer activity differences between the two strains, we next sought to predict SPRET/EiJ gene induction levels based on the motif-based linear model we built previously (Figure III-5C,D). To this end, we reconstructed the feature matrix (the input to the model) to include only the motif scores of TFBM within enhancers that are active in SPRET/EiJ, then used this new feature matrix to predict fold change in expression one hour and six hours after KLA stimulation. This new feature matrix has lower prediction errors at both one-hour post-KLA stimulation and at six hours after KLA stimulation (p-value <10^{-4} KS test) compared to using the feature matrix that includes the motif scores of TFBM within all enhancers (both active and inactive in SPRET/EiJ) (Figure III-7H, III-8C). As such, our model can incorporate the loss of TFBMs in inactivated enhancers and estimate the impact on the expression of target genes.

III.9. AP1 Family Transcription Factors Mediate the Formation of Inducible Regulatory Interactions

We previously showed that, like gene expression, chromatin accessibility and chromatin activity of cis-regulatory elements (measured by ATAC and H3K27ac respectively) have well-defined temporal patterns (Garber et al. 2012;

Donnard et al. 2018). Indeed up to 30% of regions undergo significant changes in either chromatin accessibility or H3K27ac signal after stimulation (Garber et al. 2012; Donnard et al. 2018). We found that changes in DNA accessibility and histone acetylation are not concomitant with changes in physical interactions between these regions. In fact, as others have observed in different contexts (Rubin et al. 2017; Ghavi-Helm et al. 2014; Jin et al. 2013), only a small fraction (10%) of interactions have detectable changes upon LPS stimulation (Figure III-9A top). That said, in general, stimulus-induced interactions are enriched in regions undergoing inducible H3K27 acetylation (p-value < 10^{-37}, Fisher exact test), and the associated genes tend to be induced with LPS (Figure III-9A middle). Similarly, lost interactions are enriched in genes downregulated by LPS (p-value < 0.003, Fisher exact test) (Figure III-9A bottom). These observations suggest that, although most interactions are already pre-established prior to LPS stimulation, there are subsets of genes and cis-regulatory elements that establish physical interactions only upon stimulation. The large disproportion between dynamic contacts and enhancers with LPS inducible activity suggested that connections may be established prior to enhancer activation, while enhancers are poised. One of the commonly recognized signatures of poised enhancers is the presence of H3K4me1 but lack of H3K27ac signal (Rada-Iglesias et al. 2011). For example, the Nfkbiz locus includes six poised enhancers. All of these enhancers are induced within 30 minutes of LPS stimulation. Three of these enhancers (Figure III-9B, pink boxes and loops) form interactions only after LPS

stimulation while the other three interact with the Nfkbiz promoter prior to LPS stimulation (Figure III-9B, black boxes and loops). Genome-wide we found that 80% of the induced H3K27ac peaks with stable interactions are poised enhancers, indicating that poised enhancers establish promoter interactions prior to their activation.

We next sought to investigate DNA features associated with stimulation dependent interactions. Motif enrichment analysis revealed 60 overrepresented TFBMs in dynamic interactions compared to interactions that do not show significant changes after LPS stimulation (Methods). Of these, 25 motifs are enriched exclusively in induced interactions, 9 in down-regulated interactions and the remaining 26 motifs are enriched in connections that are both induced and down-regulated. IRF, STAT, AP1, and SMAD family motifs are most strongly enriched in induced connections while KLF motifs are most strongly enriched in connections that are lost (Figure III-9C). For more than half of TFs associated with induced interactions (~54%), the TF-encoding genes are also transcriptionally induced with LPS (Figure III-9D), a ~7 fold enrichment of the expected number of induced TFs in a random sample of expressed TFs. This suggests a model where interactions that are signal-dependent require signal-dependent TFs.

TFs are typically thought to act synergistically by physically interacting with one another to regulate gene expression (Junion et al. 2012). We, therefore, investigated whether certain TF pairs mediate the formation of chromatin

interactions. To this end, we looked at pairs of motifs occurring within newly formed interacting elements such that one motif is within one interacting element and the other motif is within the other interacting pair (Figure III-9E top). We identified 25 motif pairs that are enriched (p-adjust <0.001, binomial test) (Figure III-9E) in induced interactions. Interestingly, we find that the majority of the motif pairs are for members of the activator protein 1 (AP1) transcription factor family (Figure III-9E, black boxes). Further, 5 out of 7 AP1 family members are transcriptionally induced within 1 hour of LPS treatment. Thus AP1 family heterodimers (Bejjani et al. 2019) may play a critical role in mediating context-specific inducible interactions. AP1 factors seem to mediate E-P interactions, the same TF family was previously reported to mediate long-range interaction in other cellular states as well (Qiao et al. 2015; Chavanas et al. 2008; Phanstiel et al. 2017; Vierbuchen et al. 2017).

III.5. DISCUSSION:

While it is possible to estimate the effects of coding variants on protein function, they are relatively rare and are not generally associated with common diseases (Ludwig et al. 2019; Gallagher and Chen-Plotkin 2018). In fact, most variants reside within non-coding regulatory elements (Nishizaki and Boyle 2017; Zhu, Tazearslan, and Suh 2017) and the effects of such variants on gene expression are much harder to predict. Thus, building a framework to interpret the impact of non-coding risk variants is challenging because we do not understand how regulatory elements operate to regulate gene expression.

Several reports showed that incorporating 3D interactions is critical in order to link regulatory elements to their target genes (Fulco et al., n.d.; Moore et al. 2020). However, simply knowing the target gene(s) of regulatory elements is not sufficient to estimate the impact of their loss. One way to estimate the effect of loss of a regulatory element is to build quantitative models that predict gene expression from the TFBMs they harbor. In this study, we found that SPRITE 3D interaction data greatly improves our ability to build such models and hence allowed us to test their applicability to determine the effect of genetic variability using a very well-characterized mouse system. We were surprised to find that less than 50% of putative regulatory elements have detectable contacts with each other, and that poor performance of predictive models built using distance-based gene-enhancer associations was mainly due to spurious associations. The improvement in gene expression prediction resulted mostly from the higher specificity of SPRITE-estimated E-P interactions. Together, these findings strongly suggest that accessible DNA regions, even when they are active (as indicated by the H3K27ac mark), do not necessarily exert regulatory function, and that they may instead be poised to participate in other inducible processes.

Our ability to better define regulatory interactions further helps to interpret the molecular conservation of regulatory elements. Our group and others have reported that a very small fraction of enhancers is shared across mammals (Villar et al. 2015; Donnard et al. 2018; Danko et al. 2018). Interestingly, we found that a much higher fraction (75% vs 63%) of enhancers that are shared between

mouse and human have interactions with promoters compared to enhancers that are mouse-specific. This correlation between regulatory element conservation and 3D interactions suggests that conserved enhancers may in fact be critical or less redundant regulatory elements.

The single molecule nature of SPRITE revealed transcriptional hubs and their importance. Our observations suggest that multi-way interacting complexes involving multiple enhancers and promoters may provide a mechanism to coordinate gene expression while also reducing transcriptional noise across a cell population. As SPRITE and other techniques to establish 3D chromosome conformation continue to evolve so will our understanding of enhancer biology and the models to quantify their role in gene expression. Further extensions and applications of these models to other systems will establish the practical importance of the models presented here.

III.6. METHODS:

III.6.I. Mice

All mice were housed in specific pathogen-free conditions in accordance with the Institutional Animal Care and Use Committee of the University of Massachusetts Medical School. C57BL/6 female mice were euthanized at 7-8 weeks of age to harvest bone marrow.

III.6.II. Cell Culture and Cell Lines Used

All cells were maintained at 37° C in 5% CO2 humidified incubators.

III.6.III. Mouse Bone-Marrow-Derived Dendritic Cells

Mouse dendritic cells were derived from bone marrow harvested from 6-8 week old female C57BL/6 mice. Bone marrow was then dissociated into single cells and filtered through a 70um cell strainer. The cells were then incubated with the red blood cell lysis buffer for 5 minutes. To differentiate bone marrow to dendritic cells, bone marrow cells were plated at 200,000 cells/mL in non-tissue culture treated plates. These cells were supplemented with media on day 2 and day 7. On day 5 cells were harvested and resuspended in fresh media. On day 8 all the floating cells were collected as mouse bone-marrow-derived dendritic cells. The media used for culturing and differentiating contains RPMI (Gibco) supplemented with 10% heat-inactivated FBS (Gibco), β-mercaptoethanol (50uM, Gibco), MEM non-essential amino acids (1X, Gibco), sodium pyruvate (1mM, Gibco), and GM-CSF (20 ng/ml; Miltenyi).

III.6.IV. Mouse Embryonic Stem Cells (V6.5)

V6.5 mESCs, DGCR8 knockout mESCs, and Dicer knockout mESCs were cultured in Dulbecco's Modified Eagle's Medium (Thermo Scientific) supplemented with HEPES pH 7.0, 15% fetal bovine serum (FBS), 1000 U/ml leukemia inhibitory factor (Chemicon/Millipore), 0.1 mM l-glutamine, penicillin and streptomycin, and 0.11 mM β-mercaptoethanol.

III.6.V. SPRITE

SPRITE protocol was performed as previously described in Quinodoz et al. for mouse bone-marrow-derived dendritic cells stimulated for 0 hrs, 4hrs and 24 hrs.

III.6.VI. SIP Method

III.6.VI.1. Crosslinking and Chromatin isolation

Cells were cross-linked in a single-cell suspension following steps listed in Quinodoz et al. Crosslinked cells were frozen in 5 or 10 million cell aliquots at -80C until used.

Crosslinked cell pellets (5-10 million cells) were lysed as described in Quinodoz et al with few modifications. Specifically, cell pellets were first resuspended in 1.4mL per 10M cells of Nuclear Isolation Buffer A (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 140 mM NaCl, 0.25% Triton-X, 0.5% NP-40, 10% Glycerol, 1X PIC) and incubated for 10 min on ice. Cells were pelleted at 850 g for 10 min at 4°C. The supernatant was removed, 1.4mL per 10M cells of Lysis Buffer B (10 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X Protease inhibitor cocktail) was added and incubated for 10 min on ice. Nuclei were obtained after pelleting and supernatant was removed (as above), and 550 µL of Lysis Buffer C_1 (20mM HEPES pH 7.5, 1.5mM EDTA, 100mM NaCl, 0.1% NaDOC, 0.1% Igepal CA360, 1x PIC) was added and incubated for 10 min on ice prior to sonication.

III.6.VI.2. Chromatin digestion

After nuclear isolation, chromatin was sheared via sonication of the nuclear pellet using a Branson needle-tip sonicator (3 mm diameter (1/8" Doublestep), Branson Ultrasonics 101-148-063) at 4°C for a total of 1 min at 4-5 W (pulses of 0.7 seconds on, followed by 3.3 seconds off). DNA was further digested using 0.005 - 0.01 uL of Micrococcal Nuclease (MNase) (NEB M0247S) per 10 µL of the sonicated lysate (equivalent to ~200,000 cells), in Buffer C_1 supplemented with 5mM Cacl2 at 37°C for 20 min. Concentrations of MNase were optimized to obtain DNA fragments of mean size between 150-300 bp in length. MNase activity was quenched by adding 50mM EGTA and 0.1% SDS final concentration.

III.6.VI.3. Immunoprecipitation

After digesting, the lysate is precleared with 60uL of protein A beads (at 10mg/mL concentration) per 1ug of chromatin by rotating at 4°C for at least 2 hours. Protein A beads were captured with a magnet and the lysate was transferred to a fresh tube. An equal volume of Adjustment buffer (80 mM HEPES pH 7.5, 200 mM NaCl, 1.5 mM EDTA, 50mM EGTA 1.8% Igepal CA630, 0.9 NaDOC, 0.1% SDS and 0.5mM PMSF) was added to the lysate. Next, the H3K4me3 antibody (Millipore Cat # 05-745R Lot # 288116) was added to the lysate (in this experiment a ratio of 1ug of antibody per 1ug of lysate was used) and rotated overnight at 4°C. The antibody-chromatin complexes were then captured using 25uL of blocked protein A beads (beads were blocked by rotating

overnight at 4°C in 1xPBS, 0.5% BSA, 0.5% Tween 20) and washed in ChRIPA buffer (1X PBS, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 1% Igepal CA630, 0.5% NaDOC, 0.1% SD). Chromatin is then eluted off the protein A beads using PBS-DEB buffer (1x PBS, 5 mM EDTA, 0.5% SDS, 10 mM DTT (added fresh)). 5% of the eluate was then reverse crosslinked using Proteinase K (NEB P8107S) in RNK-400 buffer (20mM Tris HCL pH8, 400mM NaCl, 10mM EDTA, 10mM EGTA, 0.5% Triton-X, 0.2% SDS) to be used for a ChIP-Seq library. The remaining eluate was then coupled to NHS beads after estimating the molarity as previously described in Quinodoz et al.

III.6.VI.4. Split-pool ligation

Split-pool ligation was also performed as previously described in Quinodoz et al. except that we used 10X lower amount of adapters in each round (4.5uM of adapters instead of 45uM).

The detailed SIP protocol is available at https://www.umassmed.edu/garberlab/protocols/

III.6.VII. SPRITE/SIP Data Processing and Cluster Generation

Barcode identification was done as previously described. The genomic reads from both the SPRITE and the SIP libraries from BMDCs and mESCs were aligned to the mm10 genome using Bowtie2(V2.3.2) with --local --trim5 11. The SAM files were converted to BAM using Samtools V1.4. BAM files are then filtered to keep only reads with all 5 barcodes identified and which are less than 2 mismatches to the reference genome. The files are then further filtered to keep

all the reads with MAPQ > 10. Clusters of interacting DNA are generated from the filtered BAM files by identifying all the reads that have the same 5 barcodes.

III.6.VIII. Viewpoint Centric Analysis

To map the enhancer-promoter interactions which are mostly short-range interactions, we used small (2-10) and medium (11-100) size SPRITE clusters (Quinodoz et al. 2018) and employed a viewpoint centric analysis (VCA) where every promoter is set as a viewpoint and all interactions occurring with it is used. To determine bins with significant interactions with a viewpoint of interest (i.e. a promoter or an enhancer), computed by fitting a negative binomial (NB) generalized linear model (GLM) that models counts based on the distance decay of the signal, and the local SPRITE read coverage (GC content did not improve the GLM fit). We next calculated the p-value for each bin pair as pij= NB(X > xij/ eij) where eij is the expected frequency of the interaction from the background model. Similar to HiC-DC, MAPS, and Fit-Hi-C, we considered bins with extremely low p-values are true positives (>=85% quantile). We removed these bins, to refit our GLM based NB model to re-calibrate the background. The p-values obtained after refitting are adjusted for FDR using the Benjamini-Hochberg procedure. Using this approach we considered 2MB region around each promoter (±2Mb) and searched for over-represented interactions with all the active promoters.

III.6.IX. Calling Interactions

Using the model described above we called interactions with all active promoters after merging all our SPRITE/SIP data from different timepoints and at each time point separately. Our calls per time point basically yielded a subset of interactions (86%) we called when all the data was merged.

III.6.X. Motif Instances

We used class A motifs from the mouse Hocomoco v11 database. All instances were detected across all ATAC peaks (promoters and enhancers) using Fimo (Grant, Bailey, and Noble 2011), with a q-value threshold of 1e-4.

III.6.XI. Random Forest

All model training and evaluation were done in R 3.5.1, using the caret (v6.0.77) (Kuhn et al., n.d.) and randomForest (v4.6.12) (Liaw, Wiener, and Others 2002) packages. For each feature set, we evaluated the accuracy of the model on the mouse data with 10-fold cross-validation. For each one of the training data in the cross-validation hyperparameters tuning was performed using 10-fold inner cross-validation with the “train” command, using the following parameters: tuneLength = 20, metric = “ROC”.

III.6.XII. Linear Regression

III.6.XII.1. K-mer identification and scoring

We used SeqGL (Setty and Leslie 2015) to identify motifs that are enriched inaccessible regions within regulatory elements of early, late, down-regulated and non-changing ATAC groups. Because each accessible site can be

of variable length (after merging from all time points), we only considered 200 bp windows centered on the summit. For the background set of regions we randomly shuffled each 200bp window within 10,000 bp on each side to pick a random region of 200 bp that matched the GC content of the test set. Finally, instead of default parameters, we used 200 groups and 30,000 features, similar to the parameters used to analyze DNase-seq data in the original SeqGL publication.

III.6.XII.2. Peak dynamics

All accessible peaks were categorized into early, late, down or non-changing as previously described (Donnard et al. 2018).

III.6.XII.3. ElasticNet regression

We used the sum of peak scores reported by SeqGL across each peak category per k-mer group as features for elastic net regression. We then used this feature matrix to predict the maximum fold change of genes at either early time point (2 hours) or late time point (6hours). The models were trained and evaluated using the caret (v6.0.77) (Kuhn et al., n.d.) and glmnet. The hyperparameters tuning was performed using 10-fold inner cross-validation with the “train” command, using the following parameters: method="repeatedcv", number=10, allowParallel = T, repeats=100, tuneGrid=expand.grid(.alpha=seq(0, 1, by = 0.05), .lambda=seq(0,10, by = 0.5)), metric = “RMSE”.

III.6.XIII. Models With Interacting Promoters

We were able to test the models with interacting promoters only at 6hrs after LPS stimulation and not at 1hr as there were only few promoter-promoter interactions of expressed promoters at this time point.

III.6.XIV. SPRET/EiJ Data Processing

We reprocessed previously published data for RNAseq and H3K27ac ChIP-Seq (Link et al. 2018) from BMDMs derived from SPRET/EiJ and C57BL/6.

The custom genome was generated for SPRET/EiJ by incorporating alleles reported in the VCF files from the Mouse genome project (version v5) (Keane et al. 2011) using EMASE (Raghupathy et al. 2018). VCF files were filtered to keep the SNPs that have pass VCF quality control and when an SNP and an indel overlapped we kept the variant with the best quality. We generated reference files using bowtie2 using the command bowtie2-build with default parameters.

The custom transcriptome was generated for SPRET/EiJ by using liftover to remap the mm10 GTF file coordinates onto the SPRET/EiJ genome created. We next generated reference files using RSEM (v1.3.0) using the command rsem-prepare-reference with the option --polyA-length 67.

III.6.XIV.1. RNAseq

We used RSEM (v1.3.0) to estimate gene expression in Transcripts per Million (TPM) with parameters --no-bam-output --bam. RSEM was configured to

use bowtie (v1.2.2). Genes with more than 15 TPMs at any time point were considered as expressed.

We used counts per gene to identify differentially expressed genes by at least -log2 fold change of 0.5 between BMDMs derived from C57BL/6 at corresponding time point whose change in expression was significant (p-adjusted < 0.05) according to the package DESeq2 (v1.10.1) (Love, Anders, and Huber 2014) in R (v3.5.1). Due to the large transcriptional changes observed in this system, we turned off the fold change shrinkage in DESeq2 with betaPrior=FALSE and we added a pseudo count of 32 to all timepoints to avoid spurious large fold change estimates from lowly abundant genes. Genes were then classified based on their response to KLA stimulation in each species (induced, downregulated or non-responsive).

III.6.XIV.2. ChIPseq

We used bowtie2 to map the reads to the custom SPRET/EiJ genome generated with default parameters. We used macs2 to call peaks with --extsize 300 option. We next merged all the peaks within 200bp of each other using bedtools slop and bedtools merge. We next built a master list of peak by merging peaks from all time points (0h,1h,6h post-KLA). The peaks were then quantified by calculating the coverage per bp at each time point and normalizing for the size of the library using DESeq2 estimatesizefactors. Then we filtered the peaks to keep all the peaks that have at least 10 normalized counts.

To get the list of peaks that are conserved between BMDMs derived from C57BL/6 and SPRET/EiJ we used liftover to map the coordinates of SPRET/EiJ onto C57BL/6 genome and all the peaks that overlapped another H3K27ac peak in C57BL/6 are considered conserved. The peaks that did not overlap the C57BL/6 H3K27ac peaks but were mappable were considered non-conserved.

Next, we compared the signal strength of conserved peaks at each time point between BMDMs derived from C57BL/6 and SPRET/EiJ mice. The peaks that have at least a log2 fold change of 2 between the two species we considered differential.

III.6.XV. Temporal Changes in Pairwise Interactions

The interactions counts for all the pairwise interactions are split based on the barcodes of 0, 4 and 24h LPS stimulation. The interactions at every time point are normalized to the depth of each library using DESeq2 (v1.10.1) in R (v3.3.5). For each interaction, we performed Fisher exact test to compare the counts at 4h or 24h against 0h to assess if the interaction is induced or downregulated or stable. We then considered any interaction to be significantly changing if the p-values obtained from Fisher-exact test is <=0.05.

III.6.XVI. Motif Enrichment in Dynamic Interactions

We performed motif enrichment in induced interactions and downregulated interactions using Fisher Exact test. The p-values obtained were corrected for multiple hypothesis testing using FDR in R (v3.3.5).

III.6.XVII. 3D Cis-Regulatory Modules

We generated all the possible combinations of motifs doublets from the significant motifs in dynamic interactions. For each pair of motifs we conditioned that they have instances in each interacting pair. We then used the FDR corrected binomial p-values test to assess their co-occurrences. For example, to compute the co-occurrences of motif x with motif y, we define the number of co-occurrences of this pair such that motif x occurs in pair1 and motif y occurs in pair2. The background probability set to the product of the probability of motif x and motif y.

III.6.XVIII. Biclustering for Finding Higher-Order Interactions

For each promoter, we built an interaction matrix where every SPRITE cluster is a row and genomic bins around the promoter are columns. We then performed biclustering on this matrix using biclust (v2.0.1) package in R (v3.5.1). To assess the random chance of getting these clusters we permuted the values of the matrix and co-clustered the rows and columns of the permuted matrix. We repeated this 1000 times and to get the p-value we used the permutation test. The active genomic bins (overlapped H3K27ac peak) that have an enriched signal in pairwise interactions are only considered.

III.6.XIX. Single-Cell Sequencing Data

Single-cell RNA-Seq data was collected using in house built inDROP for BMDMs stimulated with LPS for 0,1 and 4hours.

III.6.XIX.1. Library generation

Single cells were captured using an in house built inDrop system. Our system uses V3 beads with an 8 base UMI. First, cells were resuspended in 15% OptiPrep in 1x PBS at a density of ~80,000 cells/mL then run through the microfluidic chip of the instrument. Along with cells, V3 beads, RT mix (containing SuperScript III Reverse Transcriptase from Invitrogen 18080093) and a carrier oil (HFE 7500) were run through the microfluidic chip to capture single cells with single beads in an oil droplet. Following collection, the oligos are cleaved from the beads by a 7 minute UV light exposure, then 2 hour RT at 55C, 15-minute heat kill at 70C, followed by emulsion breaking. Next, the samples are cleaned before second strand synthesis, IVT, and RT using random hexamer primers. Then final library amplification was performed while incorporating sequencing adapters.

III.6.XIX.2. Alignment and processing

To generate fastq files we used bcl2fastq with --use-bases-mask y58n*,y*,I*,y16n* parameters. Next we extracted all the reads that contained a valid cell barcode and the unique molecular identifier (UMI) sequence had no Ns. The cell barcode and the UMI is appended to the read 1 header using custom scripts. The read 1 fastq files are then aligned to the mouse genome mm10 using TopHat (v2.1.1) (Kim et al. 2013) with default parameters. The bam files are then filtered to keep the cell barcodes that contained >=3000 reads. The filtered bam files are then processed with ESAT single-cell analysis mode (-scPrep) (Derr et

al. 2016) to generate a matrix of counts per gene per cell, using the UCSC gtf file for mouse genome built mm10. At this stage, we also extended the 3' annotations of the transcriptome file upto 1000 bases, discarded all the multi-mapped reads (-wExt 1000, -task score3p, -multimap ignore). Finally, we merged all the UMIs that have one count and are one hamming distance away from an other UMI that has two or more counts. All the scripts used for this processing are available through https://github.com/garber-lab/inDrop_Processing.

III.6.XX. Variability of Gene Expression Across Cells

The coefficient of variation (CV) is computed for all the genes at every time point. Genes are filtered to keep the ones that are detected in at least 20 cells at any time point. We then compared the CV for genes with higher-order interactions and genes without grouping genes by maximum expression and number of enhancers.

III.7. Author Contributions

Conceptualization, P.V. and M.Garber; Methodology, P.V., R.M., S.A.Q., K.G., P.M., M.Guttman, and M.Garber; Analysis, P.V. and M.Garber; Resources, M.Garber, and M.Guttman; Data Curation, P.V. and M.Garber; Writing, P.V., M.Guttman and M.Garber; Supervision, M.Garber; and Funding Acquisition, M.Garber.

III.8. Acknowledgments

We thank Oliver Rando, John Harris, Kate Fitzgerald, and members of the Garber Lab for valuable discussions and comments on the manuscript. We thank Idan Gabdank and Meenakshi Kagda for help managing our data submission to ENCODE. This project was supported by the NHGRI U01 HG007910 (M.Garber), R21CA236594 (M.Garber), NIH 4DN (U01 DA040612 and U01 HL130007) (M.Guttman), CZI Ben Barres Early Career Acceleration Award (M.Guttman), New York Stem Cell Foundation (M.Guttman), and M.Guttman is a NYSCF-Robertson Investigator.

III.9.Tables

III.10.Figures

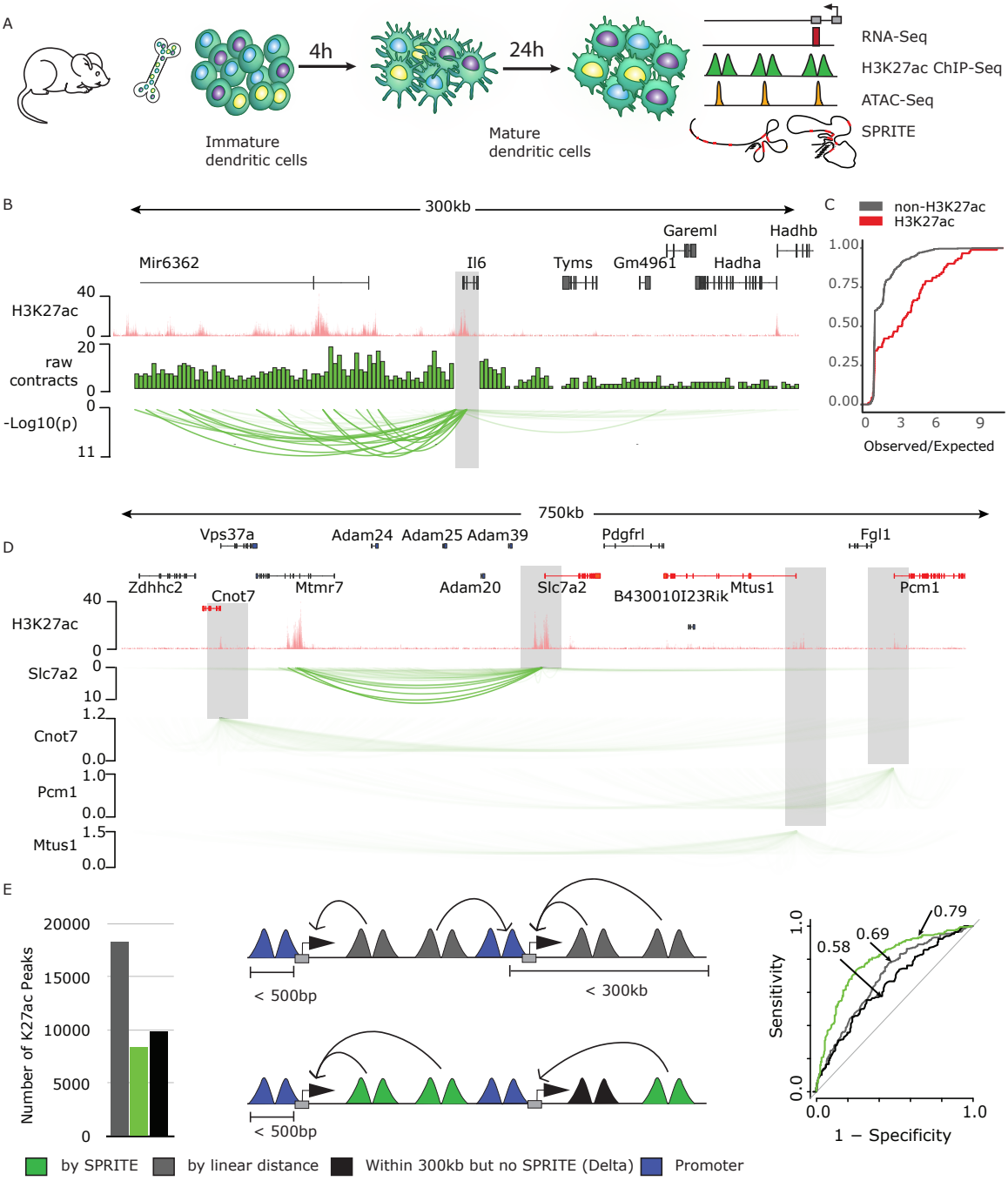
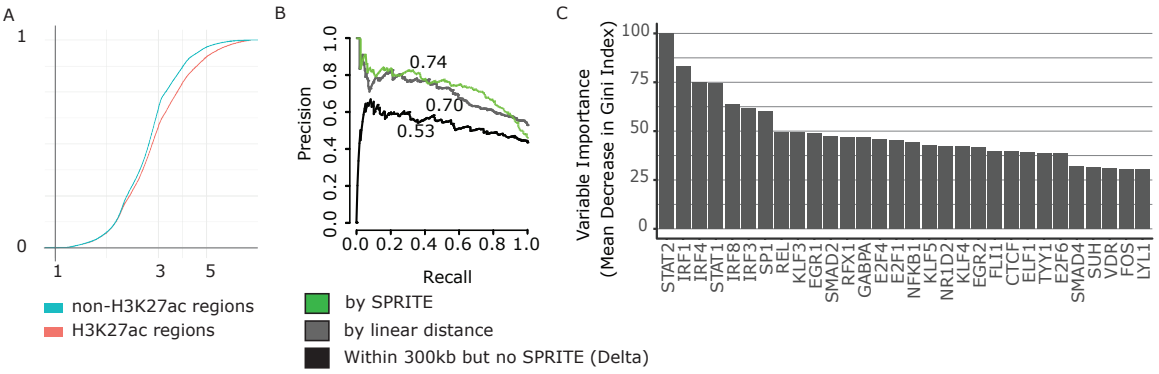


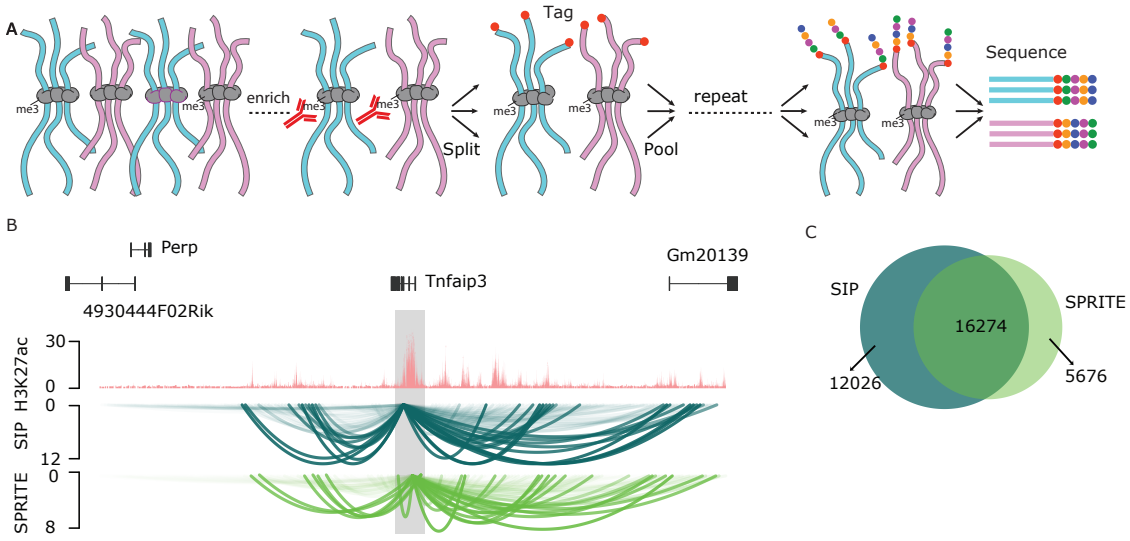
Figure III-1 | SPRITE identified enhancer promoter interactions at high resolution

(**A**) The experimental scheme used in this study. Mouse bone-marrow-derived dendritic cells are treated with LPS for 0, 4 or 24 hours. At each time point, we performed SPRITE and utilized our previously published RNA-seq, ChIP-seq, and ATAC-seq data sets. (**B**) Interaction profile of Il6 promoter. The green bars show the un-normalized contact score of every bin with the promoter. The loop plot below shows the normalized contact and heights of the loops are the -log10 p-value (p-adjusted) for each interaction. (**C**) Cumulative distribution plots of the normalized contact score of putative enhancers (red) and non-active regions (grey). (**D**) Interaction profiles of expressed genes (marked in red) in the Slc7a2 locus. The viewpoints are highlighted in grey. The heights of the loops are the -log10 p-value (p-adjusted) for the interaction. (**E**) left: Schematic depicting 1. enhancers predicted by linear proximity (grey) to promoters (purple) 2. SPRITE defined enhancers (green) 3. enhancers which have no evidence of interaction via SPRITE (delta: black). Middle: Number of enhancers associated with linear proximity, SPRITE signal or delta. Right: The ROC curve for each set of enhancers we defined.



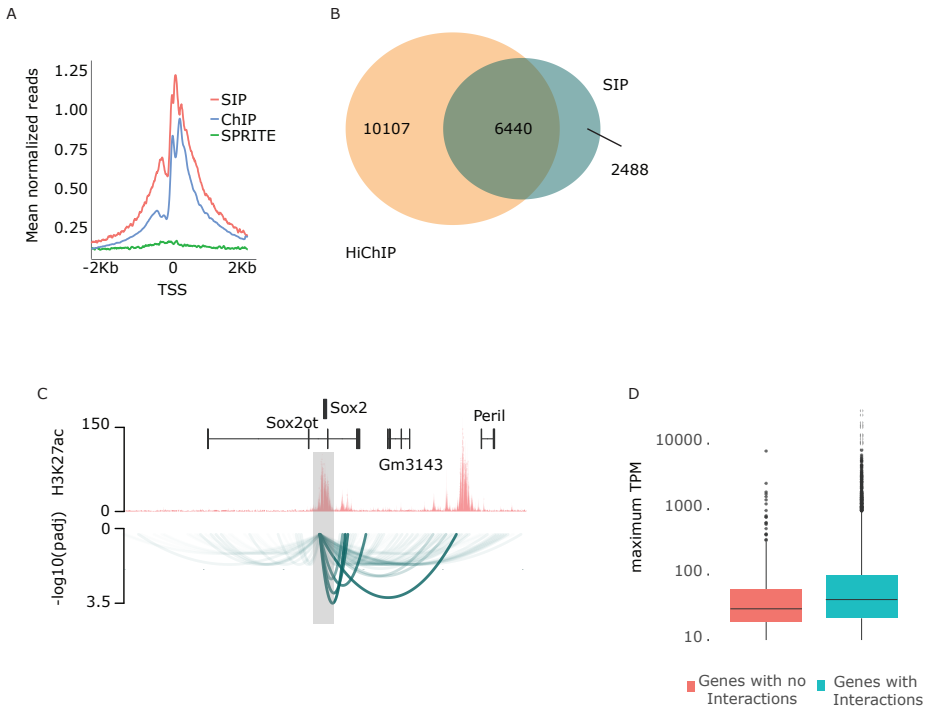
FigureIII-2 | Enrichment of putative regulatory elements in SPRITE view point centric analysis

(A) Cumulative distribution plots of the normalized contact score of putative enhancers (red) and non-active regions (grey) of all the promoters. (B) The PR-AUC curve for each set of enhancers we defined SPRITE (green), distance (grey) and delta (black). (C) Top 30 important features identified by the random forest classifier that was built using enhancers that are associated with promoters by SPRITE signal.



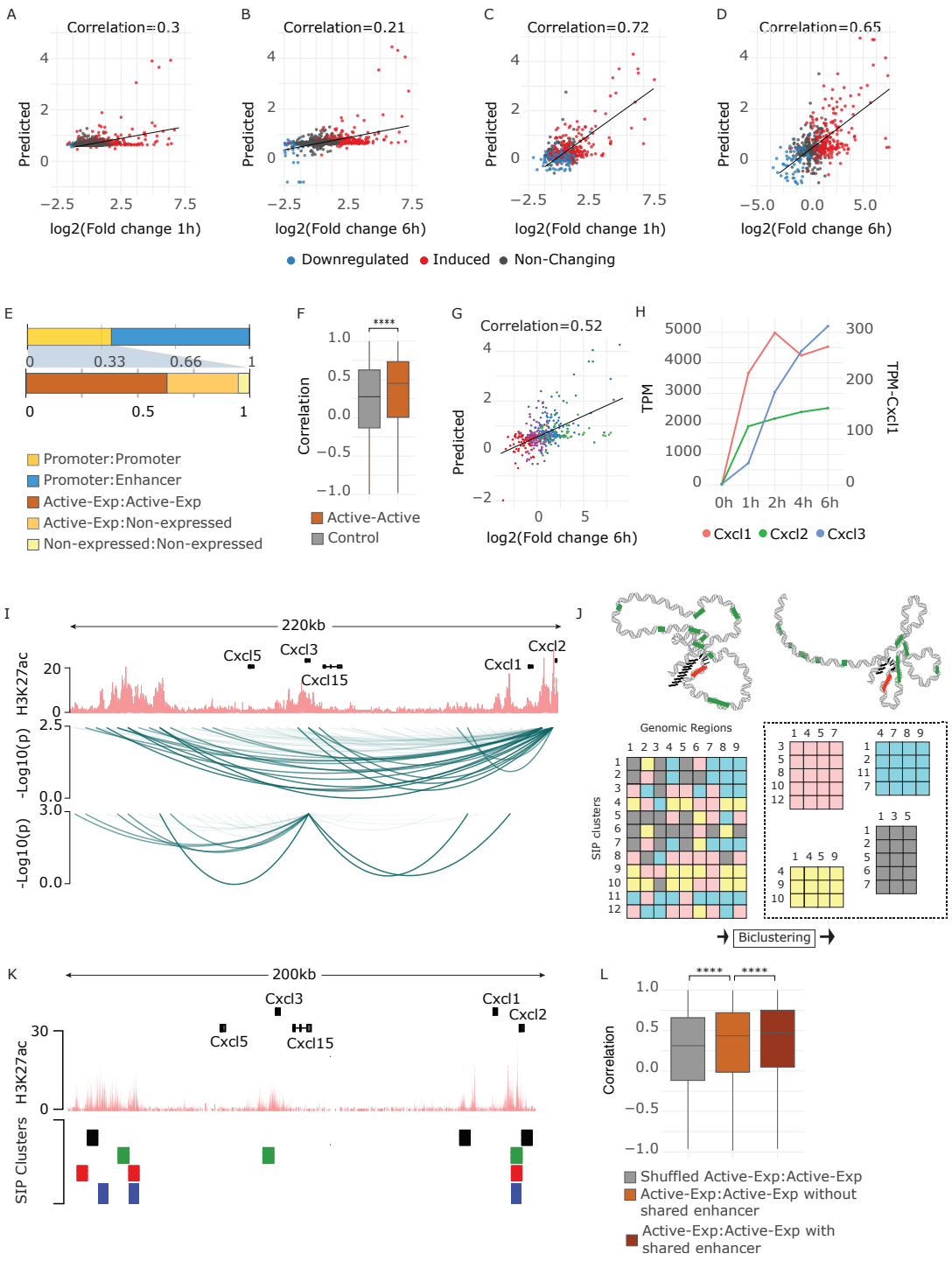
FigureIII-3 | SIP recapitulates enhancer promoter interactions identified by SPRITE

(A) Schematic of SIP workflow. Clusters of interacting DNA (pink and blue) and protein (grey) are selected with antibody (red) followed by repetitive rounds of tag extension. (B) Example locus around the Tnfaip3 gene showing the signal reproducibility between SIP -H3K4me3 and SPRITE: H3k27ac ChIP-Seq signal (pink), interactions identified by SIP-H3K4me3 (dark green) and SPRITE (light green). (C) Overlap of E-P interactions predicted using SIP-H3K4me3 and SPRITE in mDCs.



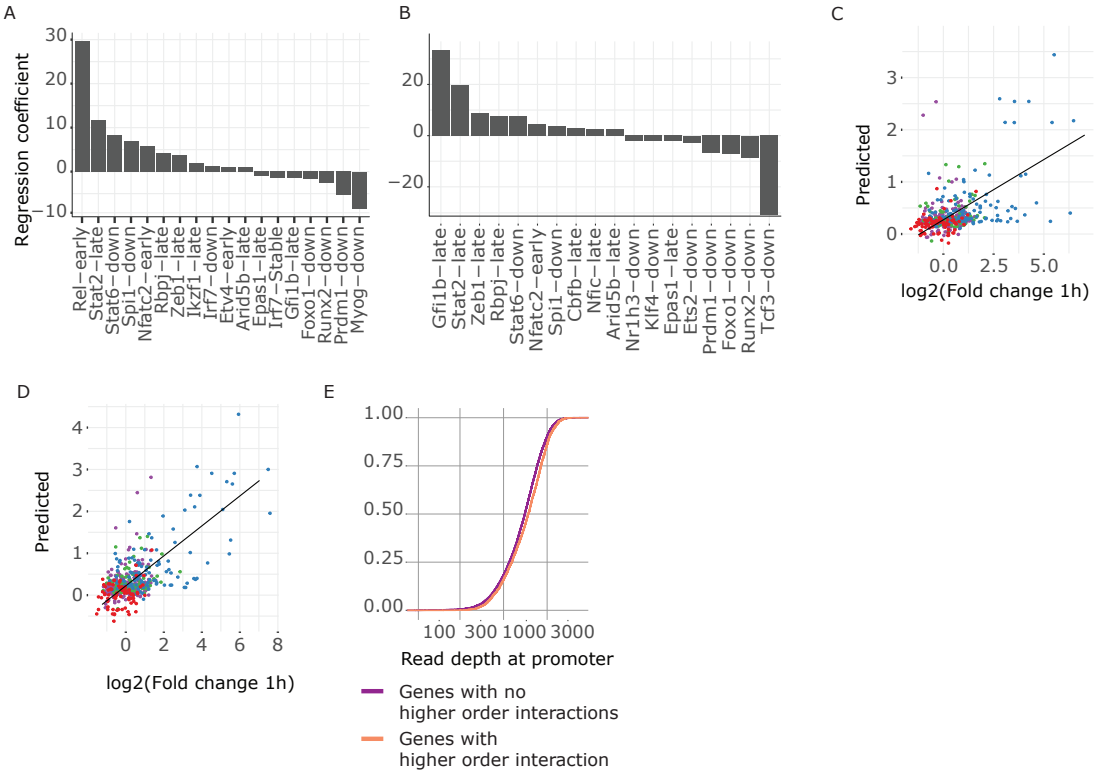
FigureIII-4 | Validation of SIP contacts in mouse embryonic stem cells

(A) Aggregate plot showing the normalized reads 2000 bp around TSS (on both sides) in mDCs for H3K4me3 ChIP-Seq (blue), SIP-H3K4me3 (red), SPRITE (green). (B) Overlap of E-P interactions predicted using SIP-H3K4me3 and HiChIP-H3K27ac in mESCs (C) The Sox2 locus in mESC: H3K27ac ChIP-Seq signal (pink) and interactions identified by SIP-H3K4me3 (dark green). (D) Distribution of maximum normalized expression for genes that have enriched interactions (blue) and do not have enriched (orange) interaction in SIP data in mDCs.


A
Correlation=0.3
Predicted
log2(Fold change 1h)
B
Correlation=0.21
log2(Fold change 6h)
C
Correlation=0.72
log2(Fold change 1h)
D
Correlation=0.65
log2(Fold change 6h)
Downregulated
Induced
Non-Changing
E
Promoter:Promoter
Promoter:Enhancer
Active-Exp:Active-Exp
Active-Exp:Non-expressed
Non-expressed:Non-expressed
F
Correlation
Active-Active
Control
G
Correlation=0.52
log2(Fold change 6h)
H
TPM
TPM-Cxcl1
0h 1h 2h 4h 6h
Cxcl1 Cxcl2 Cxcl3
I
220kb
Cxcl5
Cxcl3
Cxcl15
Cxcl2
Cxcl1
H3K27ac
-Log10(p)
J
Genomic Regions
SIP clusters
Biclustering
K
200kb
Cxcl3
Cxcl1
Cxcl2
Cxcl5
Cxcl15
H3K27ac
SIP Clusters
L
Correlation
Shuffled Active-Exp:Active-Exp
Active-Exp:Active-Exp without shared enhancer
Active-Exp:Active-Exp with shared enhancer

FigureIII-5 | Regression models predict gene-expression changes when stimulated with LPS

(A) Scatter plot of observed and predicted log2 fold change of gene-expression at one hour post LPS stimulation from a model where enhancers are associated with promoters based on linear proximity. (B) Scatter plot of observed and predicted log2 fold change of gene-expression at six hours post LPS stimulation from a model where enhancers are associated with promoters based on linear proximity. (C) Scatter plot of observed and predicted log2 fold change at one hour after LPS stimulation using a regression model that includes TFBMs from all the interactions. (D) Scatter plot of observed and predicted log2 fold change at six hours after LPS stimulation that includes TFBMs from all the interactions which include enhancer-promoter and promoter-promoter. (E) The fraction of regulatory interactions that are promoter-promoter and promoter-enhancer (top). The fraction of promoter pairs that are active-expressed:active-expressed, active-expressed:non-expressed and Non-expressed-Non-expressed (bottom). (F) Bar plots of correlation coefficinet of gene-expression across time post LPS stimulation for random promoter pairs (grey) and observed active-active promoter pairs (orange). (G) Scatter plot of observed and predicted log2 fold change at six hours after LPS stimulation using a model that includes TFBMs from active-active promoter pairs. (H) Gene expression profiles of Cxcl1-3 genes at 0,1,2,4 and 6hours post LPS stimulation. (I) H3K27ac ChIP-Seq signal across the 200kb region corresponding to the Cxcl1 gene (top). Pair-wise interactions for Cxcl2 (middle) and Cxcl3 (bottom) gene promoters. (J) Schematic showing how we identify higher-order configurations from interaction data. Two alternate conformations are shown with promoter (red), enhancers (green), and transcriptional output (black). Matrix of SIP clusters and corresponding genomic regions are deconvoluted using biclustering. (K) H3K27ac ChIP-Seq signal across 200kb region corresponding to the Cxcl1 gene (top). Higher-order interactions of SIP clusters where each configuration has at least 3 unique SIP clusters (bottom). (I) Correlation of expression across time for randomly selected gene pairs (grey), genes with interacting promoters and no shared enhancer (orange) and genes with interacting promoters and shared enhancer(s) (brown)



FigureIII-6 | Regression coefficients from the linear models

(A) Regression coefficients of top predictors of the log2 fold change of expression at one hour after LPS stimulation. (B) Regression coefficients of top predictors of the log2 fold change of expression at six hours after LPS stimulation. (C) Correlation of observed and predicted log2 fold change of gene-expression at one hour post LPS stimulation from a model where enhancers not classified based on their dynamics. (D) Correlation of observed and predicted log2 fold change of gene-expression at one-hour post LPS stimulation from a model where TFBM scores are weighted by the H3K27ac peak score. (E) Distribution of coverage for promoters that have no stable higher-order configurations (purple) and for promoters with stable higher-order configurations (orange).

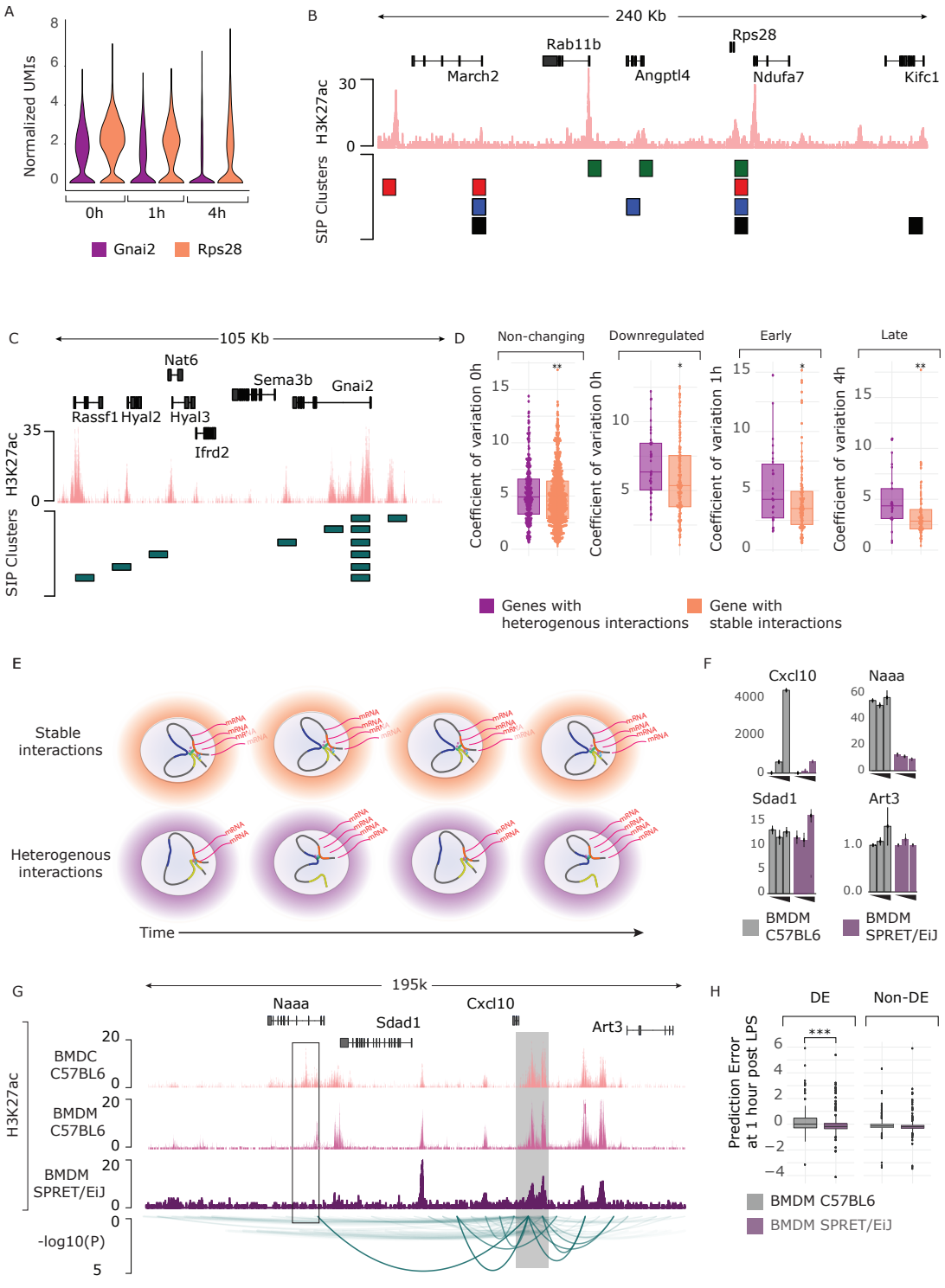


Figure III-7 | Variability in regulatory configurations predict variability in

gene expression

(A) Coefficient of variation of gene expression at 0,1 and 4 hours post LPS stimulation for Gnai2 (purple) and Rps28 (orange). (B) H3K27ac ChIP-Seq signal across 240 kb region corresponding to the Rps28 gene (top). Higher-order interactions of SIP clusters where each configuration has at least 3 unique SIP clusters (bottom) (C) H3K27ac ChIP-Seq signal across 105 kb region corresponding to the Gnai2 gene (top). Pairwise interactions of SIP clusters (bottom). (D) Coefficient of variation of expression for genes that have consistent higher-order interactions (orange) and for genes that do not have consistent higher-order interactions (purple) at 0h for non-changing genes, 0h for downregulated genes, 1h for early induced genes, 4h for late induced genes. (E) Schematic showing the correlation between chromatin interactions and gene expression at single-cell resolution. (F) Normalized expression of the genes with active promoters in the Cxcl10 locus in C57BL/6 mBMDMs (grey) and SPRET/EiJ mBMDMs (purple) at 0, 1 and 6 after KLA stimulation. (G) Example region around the Cxcl10 gene. Tracks from top to bottom show H3K27ac ChIP-Seq in C57BL/6 mBMDCs at one-hour post LPS stimulation, H3K27ac ChIP-Seq in C57BL/6 mBMDMs at one-hour post LPS stimulation, H3K27ac ChIP-Seq in SPRET/EiJ mBMDMs at one-hour post LPS stimulation, interaction-plot where the height of each loop is the -log10(p-adjusted) of the interaction. (H) Prediction errors with a regression model before (grey) and after (purple) inactivating the downregulated enhancers in SPRET/EiJ BMDMs for differentially (left) and not differentially expressed genes (right) at one-hour post LPS stimulation

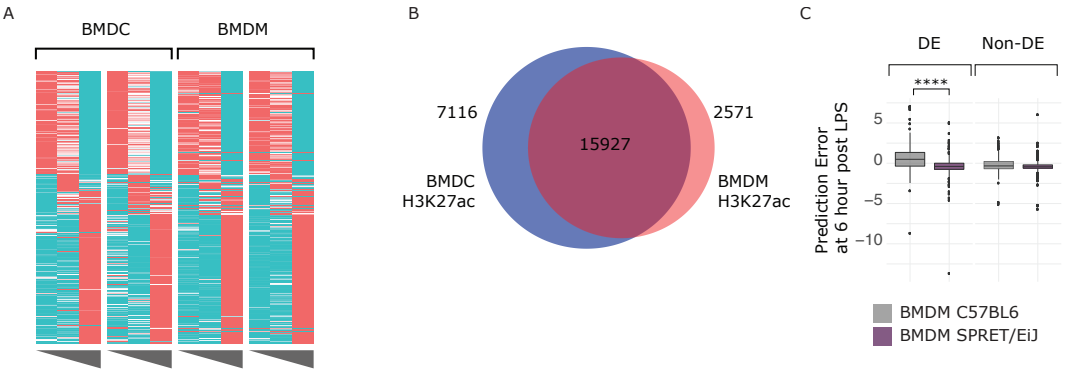


Figure III-8 | Similarity in expression and epigenetic landscape of bone marrow derived dendritic cells and macrophages

(A) Heatmap showing the normalized gene expression in response to LPS at 0, 1 and 6 hours for two replicates of BMDCs (left) and BMDMs (right). (B) Venn diagram showing the overlap of H3K27ac regions in BMDCs (blue) and BMDMs (red). (C) Prediction errors with a regression model before (grey) and after (purple) inactivating the downregulated enhancers in SPRET/EiJ mBMDMs for differentially (left) and not differentially (right) expressed genes at 6 hours post LPS stimulation.

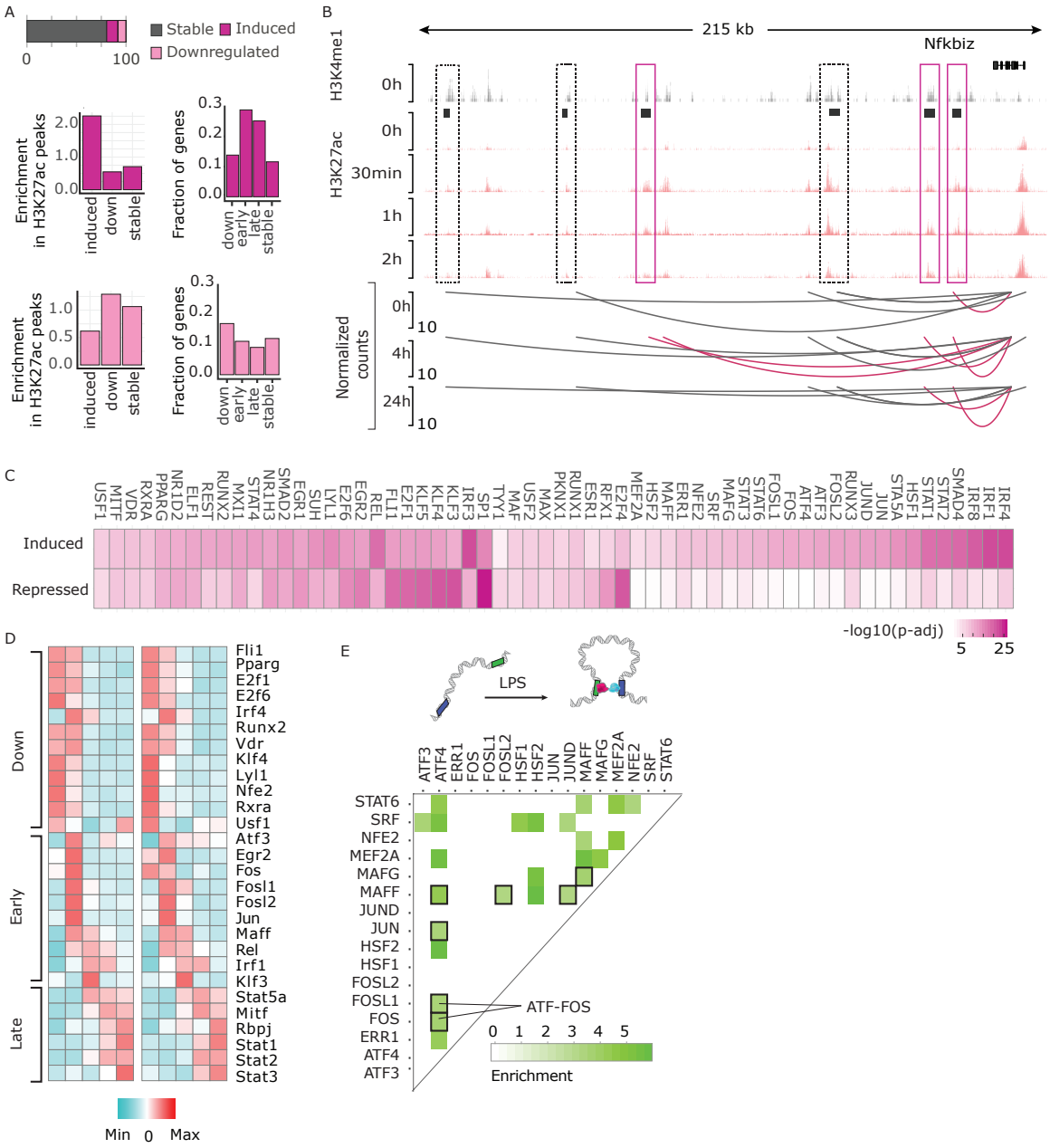


Figure III-9 | Dynamics of chromatin interactions are mediated by AP1 family transcription factors

(A) The fraction of interactions that are stable (grey), induced (dark pink) and downregulated (light pink) after stimulation BMDCs with LPS (top). Enrichment of induced, downregulated and stable H3K27ac regions (left) and fraction of downregulated, early, late and stable genes (right) in induced interactions (middle) and in downregulated interactions (bottom). (B) Example locus Nfkbiz gene (215 kb) depicting the complexity of interactions. The tracks display (top to bottom): H3K4me1 ChIP-Seq signal at 0-hour LPS stimulation, Poised enhancers, H3K27ac at 0h, 30 minutes, 1 hour and 2 hours post LPS stimulation, loops plot for normalized counts of interactions at 0h, 4h, and 24h post LPS. (C) Heatmap of -log10 p-adjusted for TFs that are enriched in induced interactions (top) and downregulated interactions (bottom). (D) Heatmap of normalized expression values for TFs that change expression when BMDCs are stimulated with LPS and are enriched in binding regulatory elements connected by changing interactions. TFs are grouped based on their expression profiles into downregulated, early or late induced. (E) Schematic illustrating the motif pairs that bind to individual regulatory elements and catalyze interaction between the elements (top). TF pairs that are significantly enriched and are colored by their -log10 p-adjusted (binomial test).

IV. CHAPTER VI: Uncovering the Short DNA Sequences Which Control the Epigenetic Landscape of Dendritic Cells Maturation

IV.1. Preface

This research chapter encompassed work performed by Shaked Afik*, Pranitha Vangala*, Elisa Donnard, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, Patrick McDonel, Jeremy Luban, Manuel Garber, Nir Yosef. The publication will be entitled “Uncovering the short DNA sequences which control the epigenetic landscape of Dendritic cells maturation”.

IV.2. Summary

Epigenetic changes are a crucial step in the cellular response to environmental stimuli and involve interactions between chromatin, non-coding DNA regions, histone modifiers, and transcription factors. Here, we present a computational pipeline to detect which DNA motifs are associated with epigenetic changes. We applied our method on human Dendritic cells stimulated with Lipopolysaccharide (LPS) which resulted in a comprehensive map of TF binding motifs that are predictive of several temporal activation patterns of regulatory regions up to 24 hours after LPS stimulation. Our results include known regulators of the LPS response, as well as TFs which interact with histone acetyltransferases and deacetylases that were previously unknown to be involved in the Dendritic cell’s response to LPS. Moreover, our computational

method is modular, generalizable and can be easily applied to study many other biological systems.

IV.3. Introduction

Changes to cell state involve activation and repression of many genes. This change is mediated by changes in chromatin accessibility and histone modifications for many putative regulatory elements that facilitate the binding of transcription factor proteins to short DNA sequence motifs within. Despite many advances in characterizing the epigenetic landscape of cells, uncovering the way all these factors interact to activate a specific process remains a challenging task.

A common way to detect binding of a given transcription factor (TF) in regulatory regions is by ChIP-seq. However, this is a laborious process that is limited to one TF per experiment. Methods to evaluate genome-wide chromatin accessibility such as ATAC-seq (Buenrostro et al. 2013) provide a genome-wide view of the accessible regions. Computational analysis of the alignment patterns of the assay within open genomic “peaks” can reveal short DNA motifs bound by a TF since the binding sites will be protected from enzymatic cleavage. Combining these alignments patterns - also known as genomic footprints - with previous knowledge of the TF binding sites can provide a simultaneous prediction for many TF bound across the genome.

Recent developments in computational pipelines have provided frameworks that can infer genomic locations bound by specific TFs from genome-

wide chromatin accessibility data (Gusmao et al. 2016; Xu et al. 2018). These methods vary in their algorithm as well as the features used for prediction but can be broadly divided into two categories: (1) motif-centric algorithms, which given a set of TF motif instances in the genome will output a per-site binding prediction (Quach and Furey 2017; Pique-Regi et al. 2011), or (2) algorithms which provide a binding prediction for the complete genome either with no DNA motif information (Li et al. 2019) or with the motif information as one of the features used for prediction (Keilwagen, Posch, and Grau 2019). The main focus of those methods has been in providing per-site prediction across the genome for a TF. Recently, several methods were developed to predict differential TF binding (Li et al. 2019; Tripodi, Allen, and Dowell 2018; Baek, Goldstein, and Hager 2017), however, they have been tested on different cell types or under different experimental conditions, which usually result in many changes to the genomic landscape.

In this work, we present a computational pipeline that extends the scope of genomic footprint algorithms to go beyond the genome-wide prediction of TF binding. We take a motif-centric approach and use supervised learning to detect changes in chromatin across regions in the genome. This generalized approach allows us to provide functional context to changes in DNA and detect which motifs are drivers for specific processes of the cell. Moreover, our pipeline does not assume prior knowledge about the structure of an “active” chromatin state (i.e. no prior assumption of a reduction in cut sites in the binding site compared to

the flanking regions), which is important as some TF do not exhibit a strong genomic footprint (Sung et al. 2014). Thus, our method can be easily extended to study the motifs which do not necessarily act as a TF binding site.

We applied our method to uncover the factors driving changes to putative active regulatory elements of human Dendritic cells in response to lipopolysaccharide (LPS). This response involves various temporal transcriptional and epigenetic changes to thousands of genes and regulatory regions in both humans and mice (Amit et al. 2009; Garber et al. 2012; Rabani et al. 2014; Donnard et al. 2018; Vandenbon et al. 2018).

Our methods discovered various DNA sequences that are predictive of epigenetic changes in the hours following LPS stimulation.

IV.4. Results

IV.4.I. Supervised Learning Approach To Detect Functional Motifs

We devised a motif-centric computational pipeline to detect which short sequence motifs are functional in a subset of genomic regions. For example, given a set of regulatory regions that are involved in the cellular response to stimulation or state change, we wish to detect the short sequence motifs within those regions that function as binding sites for transcription factors. It is important to note that the strategy we built can be readily applied to any biological system where there is a state change and it is possible to define a positive and negative set of regions based on your question of interest. A summary of the pipeline is

provided below and in Figure IV-1a, a full detailed description of the pipeline can be found in the methods section.

First, we start with a complete set TF binding motifs for TFs of interest within the accessible regions defined by ATAC-seq. Each ATAC-seq peak is assigned a positive or negative class based on the underlying question. For example, the label can be positive if this peak is a putative regulatory region in a specific cellular response. Next, for each TF binding motif, we extract the local chromatin features for each motif instance based on ATAC-seq cut sites 128bp upstream and downstream of the motif (Methods). To optimize performance we compute the cut sites only from nucleosome-free fragments (Li et al. 2019) and correct the cut sites count to account for enzymatic sequence bias (Martins et al. 2018). Instead of using the number of corrected cut sites in each base around the motif, our features are ratios between the sum cut sites of segments around the motifs at various lengths, similar to the transformation performed by msCentipede (Raj et al. 2015) (Figure IV-1b, methods). This transformation allows us to capture the spatial structure of the chromatin, without limiting the algorithm to a predefined shape. These features are then used as the input for a random forest classifier, where a motif instance is labeled as part of the positive or negative set based on the label of the ATAC-seq peak in which the motif instance is found.

A high area under the precision-recall curve (AUC PR) value indicates that the TF footprint in the positive set is distinguishable from the chromatin features

around instances in the negative set. This provides an association between this motif and the specific set of active regions. We then run this pipeline for all motifs to get a complete evaluation of the regulatory motifs which are predictive of the positive regions. A natural interpretation for a high AUC value is that changes in chromatin shape correspond to differential TF binding. However, we note that there could be other interpretations such as changes in co-binding which results in different chromatin features.

IV.4.II. Detecting Functional Motifs for TF Binding

Our pipeline is designed to detect predictive motifs within a subset of open chromatin regions. To validate the generality of our approach, we tested the ability of our computational framework to detect motifs instances within open chromatin regions that are bound by transcription factors. To that end, we ran our approach on the publicly available chromatin accessibility data (Buenrostro et al. 2013) from the GM12878 cell line. We took a set of 66 TF binding motifs, for which there exists TF ChIP-seq from the ENCODE project (ENCODE Project Consortium 2012) (Table IV-1). For each motif, our positive set was the set of motif instances within open regions that overlap a TF ChIP-seq peak, while our negative set was defined as the motif instances within open regions which do not overlap a TF ChIP-seq peak. We then computed the mean AUC PR from 5-fold cross-validation runs. Limiting our analysis to only motif instances in open chromatin regions can be challenging, as motifs from the negative set are more prone to spurious binding compared to a randomly chosen negative set of motif

instances across the genome. Despite this challenge, we are able to achieve overall high classification rates when applying a random forest algorithm to the transformed cut sites (Figure IV-1c). We are also able to achieve high classification rate, albeit slightly lower on average, when taking into account all fragment lengths, when using the cut sites prior to transformation as features, or without correcting for enzymatic bias (Figure IV-2a-c)

To further benchmark our approach, we compared the accuracy of our method to previously published algorithms for detecting TF binding (Figure IV-1c). As each method requires different input and has different parameters, we made the runs of all methods as similar as possible to our pipeline (Methods). We tested DeFCoM, an SVM based method for TF binding prediction, as well as a simple footprint depth score which describes the average cut sites in the motif compared to its surrounding region, adapted from (Baek, Goldstein, and Hager 2017). In addition, we tested Catchitt (Keilwagen, Posch, and Grau 2019), which provides a prediction for TF binding in windows of 50bp across the complete genome and was one of the winners of the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. Our approach, DeFCoM, and Catchitt all exhibit high classification rates, with no method significantly outperforming the other methods (ks test p-value > 0.84 for all pairwise comparison). All methods outperform the more simplistic footprint depth score (ks test p-value < $2*10^{-6}$). We also evaluated the performance of another genome scanning method, HINT-ATAC (Li et al. 2019), however, it achieved lower

classification rates (Figure IV-2d). This is perhaps due to the fact that the default model provided by the software was designed to work on omni-ATAC, a different experimental protocol which results in very high signal to noise ratio compared to the original ATAC-seq protocol.

IV.4.III. Uncovering the TFs That Predict Changes in the Regulatory Landscape of DC Activation

We applied our method to generate a comprehensive map of the TFs involved in temporal changes to the active regulatory landscape of human Monocyte-derived DC following LPS stimulation. To this end, we collected Monocyte-derived DCs from 5 human donors and stimulated the cells with LPS. To define the set of accessible regulatory regions, we generated ATAC-seq data before stimulation (0h) and at 30min, 2h, 4h, and 24h after stimulation (Methods). To catalog the changes in active regulatory elements post LPS we collected H3K27ac ChIP-seq data before stimulation (0h) and at 1h, 2h, 4h, 6h, 12h and 24h after stimulation. We first defined the complete set of accessible regions by finding peaks of open chromatin and combining the set of peaks from each donor in each time point to a total of 193,922 regions. Next, to generate the label the regulatory regions we computed the number of H3K27ac reads around each accessible region. Differential expression analysis (Methods) revealed 8,620 regions that show a significant change in H3K27ac signal across time. We then clustered those regions into 5 temporal activation patterns based on time of peak activation (Figure IV-3a).

We focused on the regulatory landscape of three of the temporal patterns - two early activated sets of regions, including regions which peak at 1 hour post-stimulation (“immediate-early regions”) and regions which peak at 2-4 hours post-stimulation (“early regions”), as well as the set of regions which are only activated 24 hours post-stimulation (“late-24”). We used these regions that showed a significant change in H3K27ac level post LPS as a positive set and for each case, the negative set was a randomly chosen set of motif instances from regulatory regions that show no significant change in H3K27ac levels compared to pre-stimulation at any time point (Methods). For each cluster, we ran our classification algorithm on all HOCOMOCO motifs (Kulakovskiy et al. 2018) for TFs that are expressed in at least one time point (a total of 279 motifs, Methods). For each cluster and each motif, we ran our pipeline several times: Using the ATAC cut sites from the time points of peak activation, and using the ATAC cut sites from the closest previous time point.

Our pipeline resulted in a mean AUC PR value from 5-fold cross-validation for each motif in each cluster in each time point, allowing us to rank the motifs and find the motifs which are functional in our set of activated regions. In addition, we wanted to filter motifs that did not have a better predictive value than expected by chance. To this end, we ran the pipeline on randomly assigned labels (i.e. each motif instance was randomly assigned to the positive or negative set) and computed an AUC PR (Methods). We pooled the results from 3,146

randomized runs to create a null distribution and generated empirical p-values for each AUC value in the original runs.

We detected 162 motifs which showed significant (FDR-adjusted p-value < 0.05) changes in the chromatin in at least one of the “activation time points” (i.e. 30m or 2h for the immediate-early cluster, 2h or 4h for the early cluster and 24h for the late-24h cluster) (Figure IV-3b). We also observe an increase in expression for many of the TFs associated with these significant motifs. For each set of regions, the change in expression from time point 0h to peak activation time is greater for TFs associated with significant motifs compared to TFs which binds the motifs for which we do not see significant chromatin changes (one-sided ks-test p-value < 0.02 for all sets of regions, Figure IV-3).

Upon inspecting the TFs that are predictive of the various H3K27ac temporal responses, we find that the majority (75, 47%) of the TFs are exclusively predictive for immediate-early regions. (Figure IV-3b, Figure IV-4d). Many (42/75; 56%) of these TFs also change expression in response to LPS (fold change >=2 & p-value <=0.01 by DESeq2 (Love, Huber, and Anders 2014)). This set of predictive motifs include the chromatin remodeler BPTF (Frey et al. 2017), as well as TFs that are associated with early transcriptional response to LPS such as IRF7, FOS, JUN, PU.1 (SPI1), CEBPD and STAT5 (Yamaoka et al. 1998; Garber et al. 2012; Ko, Chang, and Wang 2015; Donnard et al. 2018) (Figure IV-3c-d). In addition, we see changes in chromatin in both the immediate-early and early regions for motifs of TFs previously associated with the

transcriptional response to LPS such as REL, STAT1 and STAT2, IRF1 and IRF2 and NFKB complex (Figure IV-3c-d). Our results are also consistent with the recently published study which found PRDM1 and RARA as regulators of maturation of monocyte-derived DCs in response to HIV-1 infection (Johnson et al. 2020).

Interestingly, 33% of the motifs that are predictive in the immediate-early and early regions are also predictive in late-24h regions. These include REL, STAT1, IRF1/2, NFKB, PRDM1, and RUNX1 which were previously known to interact with histone-modifying enzymes and are LPS induced (Minnich et al. 2016; Barutcu et al. 2016; Hoogenkamp et al. 2009). These results suggest another role for the TFs involved in the early activation of the cells at a much later time point. Along with these previously known TFs, our model predicts additional factors to be associated with an active chromatin state that, to our knowledge, has not been implicated as part of the LPS response in Dendritic cells. The FOXO1 binding motif is predictive of immediate-early regions, consistent with the role of FOXO1 as a regulator of TLR4 signaling in macrophages (Fan et al. 2010). We find the motif of the Hypoxia-inducible factor 1-alpha (HIF-1A) predictive in the early and late-24h regions. HIF-1A has been previously described as having a crucial role in the inflammatory response of macrophages (Cramer et al. 2003). In the early regions, we find p63 as a predictive motif, which was shown to interact with histone deacetylases (Ramsey et al. 2011). Finally, we also observe factors that are predictive only in the late-24h regions, including

CREB1 and FOXQ1. CREB1 can interact with histone acetyltransferases and can induce an antiapoptotic survival signal in monocytes and macrophages (Yuan and Gambee 2001; Wen, Sakamoto, and Miller 2010). FOXQ1 was shown to increase pro-inflammatory potential in monocytes and involved in monocyte migration (Ovsiy et al. 2017). Of note, we also see predictive motifs who are repressors of the LPS response in macrophages such as the anti-inflammatory regulator NR3C1 (GR) (Chinenov, Gupte, and Rogatsky 2013; Chinenov et al. 2014) and NR1D1, which represses TLR4 expression and mediate temporal gating of proinflammatory cytokine responses (Fontaine et al. 2008; Gibbs et al. 2012). The association of these factors to chromatin-modifying enzymes or LPS stimulation and their role as activators or repressors in Dendritic cells needs to be experimentally determined.

To test the sensitivity of our method, we compared our results to previous methods that detect differential TF activity (Methods). First, we ran our data on the software DAStk (Tripodi, Allen, and Dowell 2018), which relies on changes in motif occurrences between two sets of regions (Figure IV-5). We also adapted the algorithm developed by Bagfoot (Baek, Goldstein, and Hager 2017), which detects TF occupancy changes based on differences in footprint depth and motif-flanking accessibility (Figure IV-6, Methods). While both methods are able to detect several of the main TFs involved in LPS stimulation, overall they show a lower sensitivity (Figures S3 and S4). This is possibly due to the low number of regulatory regions and motif instances that are used as input which limit the

sensitivity of other methods that were designed to explore genome-wide differences between different experimental conditions.

IV.4.IV. Transcription Factors Associated With H3K27ac Signal Strength

So far, our pipeline predicts TF binding in temporarily activated regions, based on a discrete classification of the H3K27ac signal. We next sought out to examine whether we can find TFs associated with the strength of the H3K27ac signal by taking advantage of the genetic variation between our samples. To this end, we called SNPs and indels using the ATAC-seq and ChIP-seq data sets generated for the 5 donors (Methods) and found 584 immediate-early regions and 438 early regions with a genetic variant in exactly one donor. For each one of those regions, we computed a z-score of the H3K27ac signal of the donor with the variant based on the H3K27ac signal distribution from the other 4 donors (Methods). Since one of the H3K27ac ChIP-seq samples at 24h had a low signal to noise ratio we excluded it from further analysis and thus decided not to test for variation between donors at 24h using only the remaining four donors. For each set of regions, we computed a motif enrichment score - a modification of the GSEA score (Subramanian et al. 2005) - to associate motifs instances with regions where a genetic variant resulted in a large change to the H3K27ac signal (Figure IV-7, Methods).

We find that at immediate-early regions, many of the motifs associated with changes to the H3K27ac signal are also predictive of the immediate-early temporal pattern during activation times, with 40% of associated motifs predictive

at 30m and 64% predictive at 2h (Figure IV-7a). Motifs associated with the H3K27ac signal include main regulators of the LPS response such as IRF1, IRF2, and RELB. Interestingly, we also observe an association between H3K27ac strength and the binding of FOXO1 as well as CXXC1, a member of the SET1 H3K4 methyltransferase complex and a regulator of macrophage phagocytosis (Lee and Skalnik 2005; Hui et al. 2018). In addition, we also find an association between H3K27ac signal and NR1D1 and MAFK which are able to interact with histone deacetylases and acetyltransferases, respectively (Yin and Lazar 2005; Hwang et al. 2013).

Surprisingly, we do not see any motifs associated with a signal strength that is also predictive of the temporal H3K27ac signal pattern of the early regions up until 4h (Figure IV-7b). During 4h we see an association for the known LPS-response regulators STAT1, STAT2, and IRF1. We also find NFIL3 associated with signal strength, which can interact with histone deacetylases (Keniry et al. 2013) as well as NFIC, which was shown to be recruited to the c-fos promoter by acetylated histones (O'Donnell, Yang, and Sharrocks 2008). We also see a few motifs associated with H3K27ac strength with a low AUC PR classification value. Those motifs include many TFs such as SP1, SP2 and NR2C2 (TR4) that can recruit histone deacetylases, and KLF16 which recruits both histone acetyltransferases and deacetylases (Doetzlhofer et al. 1999; Phan et al. 2004; Cui et al. 2011; Daftary et al. 2012). Since our classifier predicts changes in chromatin state between induced and constant regions, it will not detect TFs that

are bound genome-wide. Thus, we can conclude that while the LPS regulators are associated with H3K27ac signal changes in early activated regions, we also find SP1, SP2, KLF16 as potential factors that control acetylation at 4h post LPS-stimulation genome-wide. Another possible hypothesis is that those factors require co-binding for changes in acetylation, as it was previously shown that communication between the NFKB complex and SP1 affect histone acetylation in the promoter region of the MCP-1 gene (Boekhoudt et al. 2003). We highlight that due to the low number of regions and donors we are limited in the statistical power for this analysis, and we only present an association and not a causal effect. Nevertheless, our study suggests potential factors that control the H3K27ac signal following LPS-stimulation.

IV.5. Discussion

Computational methods to detect TF binding from open chromatin regions have provided valuable insights and is a great improvement over TF ChIP-seq as it saves time, requires fewer cells and is under more flexible experimental conditions. In this work, we aimed to expand the scope of binding prediction methods and design a pipeline built for prediction of context-dependent chromatin changes, allowing the user to detect changes only in a subset of genomic regions of interest. This is designed as a highly modular framework that can be applied to any system with state change to understand the predictability of TF footprint to either gene expression changes, chromatin state changes or any other label as long as the user can define a positive set of activated regions.

Our pipeline is highly modular, can be used in conjunction with several different software achieving high classification results, allowing for many researchers to adapt easily with their existing pipelines. In addition, our method has no prior assumption about the expected shape of the chromatin around motifs from the positive set, thus this framework is easily expandable to test the importance of short regulatory motifs which are not known TF binding motifs. It should be noted that according to the motif classification suggested by HOCOMOCO, certain motifs are low confidence and can be found only in a small number of regions. We need to be cautious and not over-interpret the results from these motifs as they can be due to technical artifacts.

Changes to the epigenome landscape are an important component of the cellular response of DCs to pathogens (Boukhaled et al. 2019). Here, we aim to gain a greater understanding of the factors involved in histone modifications during DC maturation in response to LPS up to 24hrs post-stimulation. Applying our framework we identified many TFs that could be potentially be involved with chromatin-modifying enzymes to establish signatures of active chromatin (H3K27ac). Of the TFs we predict to be important for activating regulatory regions, many are previously shown to interact with chromatin-modifying enzymes, some interact with acetyltransferases while some interact with deacetylases. We highlight that our current data does not allow us to claim any causal relations between the TFs and the active regions. Thus, we cannot determine which of the TFs actively modify the chromatin and which bind to these

regions because of their active state. Determining the exact interactions as well as which chromatin-modifying enzyme the TFs interact with has to be done in future validations. Nevertheless, our work resulted in a valuable map of temporal TF-DNA interactions of the human response to pathogens and provides an easily extendable framework to be used to answer many other biological questions.

IV.6. Methods

IV.6.I. Human Subjects:

Anonymous, healthy donor leukopaks (New York Biologics, Southampton, NY), were used in accordance with UMMS-IRB protocol ID #H00004971.

IV.6.II. Cell Culture:

All cells were maintained at 37° C in 5% CO2 humidified incubators.

IV.6.II.1.Human Monocyte-Derived Dendritic Cells:

Human dendritic cells were derived from peripheral blood mononuclear cells (PBMCs) isolated from de-identified, healthy donor leukopaks (New York Biologics, Southampton, NY), in accordance with UMMS-IRB protocol ID #H00004971. Mononuclear leukocytes were isolated by gradient centrifugation on Histopaque-1077 (Sigma-Aldrich, St. Louis, MO). CD14+ mononuclear cells were enriched via positive selection using anti-CD14 antibody MicroBead conjugates (Miltenyi, San Diego, CA), according to the manufacturer's protocol. CD14+ cells were then plated at a density of 1 to 2 x 106 cells/ml in RPMI-1640 supplemented with 5% heat-inactivated human AB+ serum (Omega Scientific, Tarzana, CA), 20 mM L-glutamine (ThermoFisher, Waltham, MA), 25 mM HEPES

pH 7.2 (Sigma-Aldrich), 1 mM sodium pyruvate (ThermoFisher), and 1 x MEM non-essential amino acids (ThermoFisher). Differentiation of the CD14+ monocytes into dendritic cells (human DCs) was promoted by addition of recombinant human GM-CSF and human IL-4; cytokines were produced from HEK293 cells stably transduced with pAIP-hGMCSF-co or pAIP-hIL4-co, respectively, as previously described (Reinhard et al. 2014), with each cytokine supernatant added at a dilution of 1:100.

IV.6.III. Library Preparation and Sequencing

IV.6.III.1. ATAC-Seq

For each time point, 5 x 105 scraped DC's were collected by centrifugation 500 x g for 5 min. and lysed for ATAC-seq following the protocol described in (Buenrostro et al. 2015). Each sample was tagmented using 12.5 ul Nextera TDE-1 transposase (Illumina) for 30 minutes at 37, then quenched by the addition of 5 volumes DNA Binding Buffer (Zymo Research) and cleaned using Zymo Research DNA Clean and Concentrator-5 columns according to the supplied protocol. Tagmented DNA was PCR-amplified using indexed primers as described in (Buenrostro et al. 2015), using total cycle numbers for enrichment as determined empirically by qPCR to minimize PCR duplicates. The resulting libraries were purified twice by Zymo Research DNA Clean and Concentrator-5 columns using a ratio of 5:1 DNA Binding Buffer: Sample, and quantified by Qubit HS-DNA Assay (Thermo Fisher Scientific) and Bioanalyzer High-Sensitivity DNA

(Agilent Technologies). Final ATAC-seq libraries were pooled (equimolar) and sequenced on an Illumina Nextseq 500.

IV.6.III.2. ChIP-Seq

Harvest and Formaldehyde crosslinking. For each timepoint and donor, 5-7 x 106 unstimulated or LPS-stimulated dendritic were harvested by scraping in medium and centrifugation at 500 x g for 5 minutes. Each cell pellet was washed once with 2 mL PBS and gentle flicking of the tube, followed by centrifugation at 500 x g for 5 min. Cells were uniformly resuspended in 1 mL 1X Fixing Buffer A from the Covaris tru-ChIP Chromatin Shearing and Reagent Kit and fixed by adding 1 mL 2% methanol-free formaldehyde (Thermo Fisher Scientific) diluted in 1X Fixing Buffer A (1% formaldehyde final, 2.5-3.5x106 cells/mL) and rotated end-over-end for 5 min. at room temperature. Fixation was quenched by adding 240 mL Quenching Buffer E (Covaris tru-ChIP kit) and rotating for an additional 5 min. Purified BSA was then added to 0.5% w/v final to prevent cell adherence to the tube, and crosslinked cells were harvested by centrifugation, 500 x g for 5 min. at 4oC. Crosslinked cells were washed twice in 2 mL ice-cold PBS + 0.5% BSA with centrifugation as above, and aliquoted evenly into 3 fresh 1.5 mL tubes during the second wash. Cells were finally pelleted by centrifugation at 16,000 x g, flash-frozen as dry pellets in liquid nitrogen, and stored at -80oC.

Lysis, Shearing, and Quantification. Individual crosslinked cell pellets (1.5-2 x 106 cells each) were lysed according to the Covaris tru-ChIP Chromatin Shearing and Reagent Kit instructions. Following lysis, nuclei were resuspended

in 130 mL ice-cold Shearing Buffer D3 and transferred to 1.5 mL BioRupter Pico Microtubes (Diagenode) on ice. Chromatin was sheared to uniform fragment lengths (150-400 bp) by sonication at 4oC in a BioRupter Pico (Diagenode) set to 6 cycles of 30s ON and 30s OFF. Sheared chromatin was diluted in 10 volumes of ChRIPA buffer (1X PBS, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5% sodium deoxycholate, 1% Igepal CA-630, 0.1% SDS, 1X Roche cOmplete Protease Inhibitor Cocktail) and insoluble material was removed by centrifugation >15,000 x g for 10 minutes. Lysate was pre-cleared against 60 mL Dynabeads Protein A (Thermo Fisher Scientific) per 106 cells for 2h at 4oC with end-over-end rotation followed by two rounds of magnetic bead removal and transfer to fresh tubes. 2% of pre-cleared lysate was removed for DNA quantification and the remaining lysate was either flash-frozen in liquid nitrogen and stored at -80oC, or stored overnight at 4oC for use in immunoprecipitation. For quantification, 2% pre-cleared lysate was treated with 10 mg RNase A (Thermo Fisher Scientific) for 30 min. at 37oC, followed by addition of 100 mg Proteinase K (New England Biolabs) and crosslink reversal overnight at 65oC. DNA was purified using DNA Clean and Concentrator-5 columns (Zymo Research). Average sheared DNA fragment sizes were determined by agarose gel and chromatin yield was estimated by Qubit HS-DNA Assay. 50-100 ng purified DNA was saved as Input.

Chromatin Immunoprecipitation. Antibodies used for ChIP were rabbit anti-H3K27ac (Diagenode C15410196) and rabbit anti-H3K4me3 (EMD Millipore

05-745R). 1 mg antibody was added to 0.5 mg (anti-H3K27ac) or 1 mg (anti-H3K4me3) pre-cleared crosslinked lysate and incubated overnight with continuous mixing at 4oC. IgG/chromatin complexes were captured for 1h at room temperature on 25 mL Dynabeads Protein A that were pre-blocked for at least 1h with Blocking Buffer (1X PBS, 0.5% BSA, 0.5% Tween-20). Complexed beads were washed 5 times with ice-cold ChRIPA Buffer, twice with room temperature RIPA-500 Buffer (10 mM Tris pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS), twice with ice-cold LiCl Wash Buffer (10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal CA-630, 0.5% sodium deoxycholate), and twice with ice-cold TE buffer. Each chromatin sample was eluted from beads using 50 ul Direct Elution Buffer (10 mM Tris pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.5% SDS) and supplemented with 20 mg RNase A, incubating for 30 min. at 37oC. 20 mg glycogen was added to each bead/eluate suspension, and crosslinks were reversed by the addition of 50 mg Proteinase K and incubation at 37oC for an additional 2h, followed by overnight at 65oC. Dynabeads were removed by magnet capture, and the supernatant was mixed thoroughly with 2.3 volumes of Agencourt AMPure XP (Beckman Coulter) bead suspension and incubated for 10 minutes at room temperature prior to bead capture and washing. Purified DNA was eluted in 10 mM Tris pH 8.0.

Library Preparation and Sequencing. Sequencing libraries were prepared from half of each ChIP sample and 50 ng Input DNA using the Ovation

Ultralow System V2 kit (NuGEN) according to supplier's instructions, with the total numbers of enrichment PCR cycles determined empirically for each sample by qPCR to minimize PCR duplication rates. Barcoded libraries were quantified using Qubit HS-DNA Assay, qualified using Agilent Bioanalyzer High-Sensitivity DNA, and pooled for sequencing on Illumina Nextseq 500.

IV.6.IV. Alignment and Processing of Reads

Genome reference:

All the data generated and used for this paper is aligned to human reference genome hg19

ATAC-Seq: Paired-end reads were trimmed to remove adapter sequence using Cutadapt version 1.3, and then aligned to reference genome hg19 with Bowtie2, version 2.1.0, parameter –X 2000. The alignments were then filtered using Samtools (Li et al., 2009), version 0.0.19, to remove (i) PCR duplicates, as identified by Picard's MarkDuplicates, and (ii) aligned reads with mapping quality below 4. While the reads were aligned as paired-end to optimize the alignment accuracy, the alignments were then further processed as if they were aligned single-end sequence data, so that each aligned read corresponded to a Tn5 cut-site.

Peak Calling: Each aligned read was first trimmed to the 9-bases at the 5'-end, the region where the Tn5 transposase cuts the DNA, and then extended 10-bases upstream and down, for smoothing. Peaks were called using these

adjusted 29-base aligned reads with MACS2 (Zhang et al., 2008)], parameters --bw 29 --tsize 29 and --qvalue 0.0001.

Quality Control: Following the standard practice (Buenrostro et al., 2015), for each sample, we examined the fragment length distribution, as well as a comparison of the aggregate nucleosome signal to the aggregate nucleosome-free signal over transcription start sites for those genes found to be expressed for at least one time point in our RNA-Seq time series. Signal-to-noise ratios were computed for the peaks as f/(1 –f) where f is the fraction of reads overlapping peaks.

ChIP-Seq: Paired-end reads were trimmed to remove sequencing adapters and leading and trailing bases with quality scores less than 5. Reads that were longer than 36 bases after trimming were kept for further analysis. The reads were then aligned to human reference genome hg19 using Bowtie2 with options -k 1 --un-conc to filter out reads that map to multiple locations in the genome and that align un-concordantly. Duplicated reads were filtered out using picard-tools-1.131 MarkDuplicates function. Peaks were then called using MACS2 with --bw=230 --tsize=75 and --qvalue 0.0001.

IV.6.V. ATAC Normalization

To include only nucleosome-free fragments, we filtered out fragments longer than 180 bp with the alignmentSieve command from DeepTools (Ramírez et al. 2016). Next, to correct sequence bias due to enzymatic sequence preferences we ran seqOutBias (Martins et al. 2018) to get a per-base estimate

of read counts. We ran the correction on the length-filtered reads, performing the correction on each strand separately with the parameters “--read-size=35 --shift-counts” and the k-mer mask as recommended by the seqOutBias tutorial:

plus_mask=NXNXXXCXXNNXNNNXXN

minus_mask=NXXNNNXNNXXCXXXNXN

IV.6.VI. Classifying ATAC Peaks Based on K27 Signal

For each ATAC peak, we extracted the number of H3K27ac ChIP-seq reads that align within the peak in each sample, extended by 1000bp on both sides to include bordering histones and merging peaks that were overlapping due to that extension. We removed one sample (donor F33 at 24h after LPS stimulation) due to very low read alignment across all peaks.

To detect peaks with temporal changes, we performed pairwise differential expression using DEseq2 (Love, Huber, and Anders 2014), by comparing all the time points with time point 0h (prior to LPS stimulation) and adding the batch as a covariate to the model. Each peak that had an adjusted p-value < 0.05 and an absolute log fold change > 2 were considered as temporal peaks. In addition, we also searched for peaks which show a continuous temporal change with ImpulseDE2 (Fischer, Theis, and Yosef 2018), and added to our set of temporal peaks regions which ImpulseDE2 adjusted p-value <= 0.05. Then, we used k-means clustering with k = 5 to cluster the temporal regions into different temporal clusters. The rest of the regions were classified as constant regions, excluding

regions with a low number of aligned ChIP-seq reads (Total normalized count across all samples < 30).

IV.6.VII. RNA-Seq Analysis

Reads were aligned to the transcriptome with RSEM. To detect differentially expressed genes we ran DEseq2 for each time point. We took only expressed genes (genes with an average TPM of 10 in at least one time point) as our set of expressed genes.

IV.6.VIII. Motif Scanning

The full set of TF binding motif was downloaded from HOCOMOCO v11 (Kulakovskiy et al. 2018). We focused only on motifs of TFs which are expressed in our system, leaving a total of 279 motifs for TFs with at least an average of 10 TPM at any of the time points. We used PWMscan (Ambrosini, Groux, and Bucher 2018) to find motif instances across the genome with the "pwm_mscan_wrapper" command and parameter -e 0.00001

IV.6.IX. A Supervised Learning Algorithm for Detecting K27 Patterns

We repeated the following pipeline for each one of the 278 TF binding motifs:

IV.6.IX.1. Labels

For a given temporal cluster, we define our positive set by finding all the motif instances that fall within that set of ATAC peaks with BedTools (Quinlan and Hall 2010). The negative set is then selected out of all motif instances that fall

within the set of constant peaks, downsampled so that the two sets are of equal size. We only ran our pipeline on motifs that had at least 10 motif instances within the positive set.

IV.6.IX.2. Features

We use the ATAC-seq data as the features for our classifier. Every time point is separate, thus for a given motif and a given positive set, we run several classifiers, one for each ATAC-seq time point. We compute the normalized number of ATAC cut sites in each base at a region of 256 bp centered on the motif, where we combine the normalized cut site count from both strands oriented around the motif (i.e. cut site count 5 bp downstream on the plus strand was combined with the cut site count 5 bp upstream on the negative strand).

We represented the cut site features as ratios between regions around the motif, at different levels. The first level is the sum of cut sites around the motif, corrected for library size with the sample-specific DEseq scaling factor. The second level includes the sum of reads of the first half of the window (positions 1-128) divided by the total number of cut sites. The third level includes the first quarter compared to the first half and third quarter compared to the second half. We continue in a similar fashion until the last level in which each odd-numbered position is divided by the sum of cut sites in its position and the subsequent one. The total number of features is identical to the number of cut sites. Each donor was considered a separate, thus each motif instance translates into 5 samples in the classifier (one for each donor).

IV.6.IX.3. Classification

We ran a random forest classifier using 5-fold cross-validation with the R caret package (Kuhn 2008). We divide our samples into training and testing based on genomic position, thus for each motif instance data from all 5 donors is either all part of the train set or all part of the test set. In each run, we use the inner 5-fold CV to tune hyper-parameters and apply the model on the test set to compute the area under the curve of the PR curve with the PRROC package (Grau, Grosse, and Keilwagen 2015). The final AUC value is determined as the average of the 5 runs.

IV.7. Testing the Significance of AUC PR Values

For a given condition (i.e. a combination of ATAC time point and cluster label) we collected the set of motifs that achieve an AUC value of at least 0.5. To generate a random distribution of AUC PR values, we randomly select a motif from that set and run the same pipeline as before, except after downsampling we randomly re-assign each motif instance to the positive or negative set, keeping all biological repeats as all positive or all negative. We repeated this process for 8 different conditions (for the labels of peak 1h, peak 2h, and peak 24h, collecting both the peak time points and the previous time point) 200-400 times. As each condition showed a similar distribution of randomized AUC PR values (t-test p-value > 0.4113 for all pairwise comparisons), we collected all results across all conditions to form the null distribution with a total of 3,146 AUC PR values. We

then computed an empirical p-value for each AUC PR value for our non-random runs and performed FDR correction for all motifs in each condition separately.

IV.8. Validation of Pipeline on GM12878 Data

IV.8.I. Data Preprocessing

We downloaded previously published ATAC-seq data on GM12878 (Buenrostro et al. 2013). Reads were aligned to hg19 using bowtie2 (Langmead and Salzberg 2012). We removed low-quality alignments (MAPQ < 10) and reads without a unique alignment, as well as discordant reads and reads mapping to chrM or the ENCODE “blacklist” regions.

For peak calling, reads aligned to the positive strand were shifted +4bp, and reads aligning to the negative strand were shifted -5bp. We called peaks using MACS2 on the cut sites, merging peaks that were less than 10bp apart, leaving a total of 203,977 peaks.

For footprint method evaluation, we removed reads with fragment length > 180bp for all methods except HINT-ATAC, as HINT incorporates the fragment length as part of its model. To account for sequence bias we normalized the data with seqOutBias as described in their tutorial. TF ChIP-seq peak files for GM12878 were downloaded from the ENCODE portal (Davis et al. 2018). A full list of the TF bed files used is provided at supplementary table 1

IV.8.II. Label

For each TF binding motif, our positive set was defined as the set of motif instances within an ATAC peak that overlap the corresponding TF ChIP-seq

peak, while the negative set is are the motif instances within ATAC peak that do not overlap the TF ChIP-seq peak. We downsampled the set to have equal sizes and removed motifs that had less than 100 total samples after downsampling. The rest of our pipeline was performed as described above for H3K27ac.

IV.9. Footprint Depth Score

The footprint depth score was adapted from (Baek, Goldstein, and Hager 2017). In each motif instance, we define the footprint depth score as the 10% trimmed mean normalized cut site within the motif (extended 2 bp from the motif boundary). From this value, we subtract the mean normalized cut site in the regions flanking the motif, up to a window of 256bp (same window used for the random forest classifier). We multiplied the score in -1 so that a more positive score will be associated with greater footprint depth. Using that score, AUC PR was computed to each one of the five testing data used for the random forest classifier.

IV.10. DeFCoM

DeFCoM requires a BAM file as input, thus we ran DeFCoM on the processed BAM file after removing long fragments and read shifting, with the default parameters described in the example config file from the DeFCoM website. In each iteration of the 5-fold cross-validation we ran DeFCoM with the same test and train motif instances as the random forest classifier.

IV.11. Catchitt

Labels for each 50bp window across the genome was computed with Catchitt's "labels" commend, with the encode ChIP-seq peak file as input. Chromatin accessibility was computed with the "access" command, providing the processed BAM file as input. The "motif" command provided motif scores for each window. Since training and testing are performed on entire chromosomes we implemented a greedy algorithm to ensure a balanced 5-fold training set. We ranked the chromosomes based on the number of ChIP-seq peaks found in it from highest to lowest. We then sorted the chromosomes into five sets, each time adding the remaining chromosome with the highest number of ChIP-seq peaks into the bin with the lowest total number of ChIP-seq peaks. In each iteration, one of the sets was used for training while another set was used for testing. For the testing chromosomes, we computed the AUC PR of windows that overlap motif instances that fall within an ATAC-seq peak, with the instance label determined by the ChIP-seq and downsampling the positive and negative set to be of equal sizes. In case a motif instance spanned the edges of two windows, the score of the instance was the mean of the score for the two windows.

IV.12. HINT-ATAC

We ran HINT-ATAC on the full set of peaks and on all aligned reads with the following parameters: "--atac-seq --paired-end --organism=hg19".

To compute AUC values, for each set of test data used by the random forest classifier we assigned each motif the HINT score that overlaps it, or zero if

it didn't overlap any HINT footprint. We used those values to compute the AUC score for each test data.

IV.13. DAStk

For each set of regions, we ran DAStk using default parameters, comparing the set of induced regions to the set of regions classified as constant. MD score was computed as the difference in the MD scores between the induced and the constant regions from the output to the "differential_md_score" command.

IV.14. BagFoot

We adapted the Bagfoot algorithm as described by (Baek, Goldstein, and Hager 2017) to run on our normalized data. For each motif instance, we computed the footprint depth score and the flanking accessibility score. The footprint depth score is calculated as above, except we take a window of 200bp. The flanking accessibility score is the mean normalized cut site count of the 200bp centered around the motif. The footprint depth difference is taken as the difference between the mean footprint depth score of motif instances in the induced region and the mean footprint depth score of motif instances in the constant regions, and similarly for the normalized cut count difference. The Bagplot and p-value and adjusted p-value computation were done with the code of the gen_bagplot_chisq function, with minor modification to work on our input data.

IV.15. SNP and Variant Calling

We used GATK with the steps described in the best practices guide published by the GATK developers: https://software.broadinstitute.org/gatk/best-practices. We performed read grouping, base quality score recalibration (BQSR) on the BAM files for which we used the default parameters dbSNP-147 VCF file. And the raw variants identified by the genotyping tool were recalibrated using dbSNP-147 VCF file and default parameters. The variants are then refined using VariantFiltration, with parameters QualByDepth, FisherStrand, StrandOddsRatio, RMSMappingQuality, MQRankSum, and ReadPosRankSum were <2.0, <40.0, >60.0, >3.0, <12.5 and <−8.0 as recommended, respectively.

IV.16. Association Between H3K27ac Signal and Motif Abundance

For this analysis, we only considered regions classified as immediate-early or early regions with a variant in only one donor. For each peak, we computed the z-score of the normalized H3K27ac signal for the donor with the variant based on the mean and standard deviation of the signal from the other 4 donors. For a given set of regions and a given time point, we performed an enrichment test to test whether a motif is enriched in peaks with high z-scores: for each motif, we counted the number of motif instances in each region. We computed the enrichment score as described in (Subramanian et al. 2005), where the number of motif instances in the peak was used as the magnitude of increment in each step and the absolute value of the z-score was the weight of

each region, normalized to sum up to one. We used the absolute z-score since we wanted to test the association between a TF motif and the magnitude of the effect of a variant on the H3K27ac signal. However, we don't observe a significant change in the results when taking the signed z-score (Figure IV-8) or when the magnitude of increment was one if the region had a motif instance (instead of the number of motif instances, Figure IV-9). For the significance of the enrichment score, we shuffled the z-scores between the peaks 10,000 times and computed an empirical p-value.

IV.17. Tables

IV.18. Figures

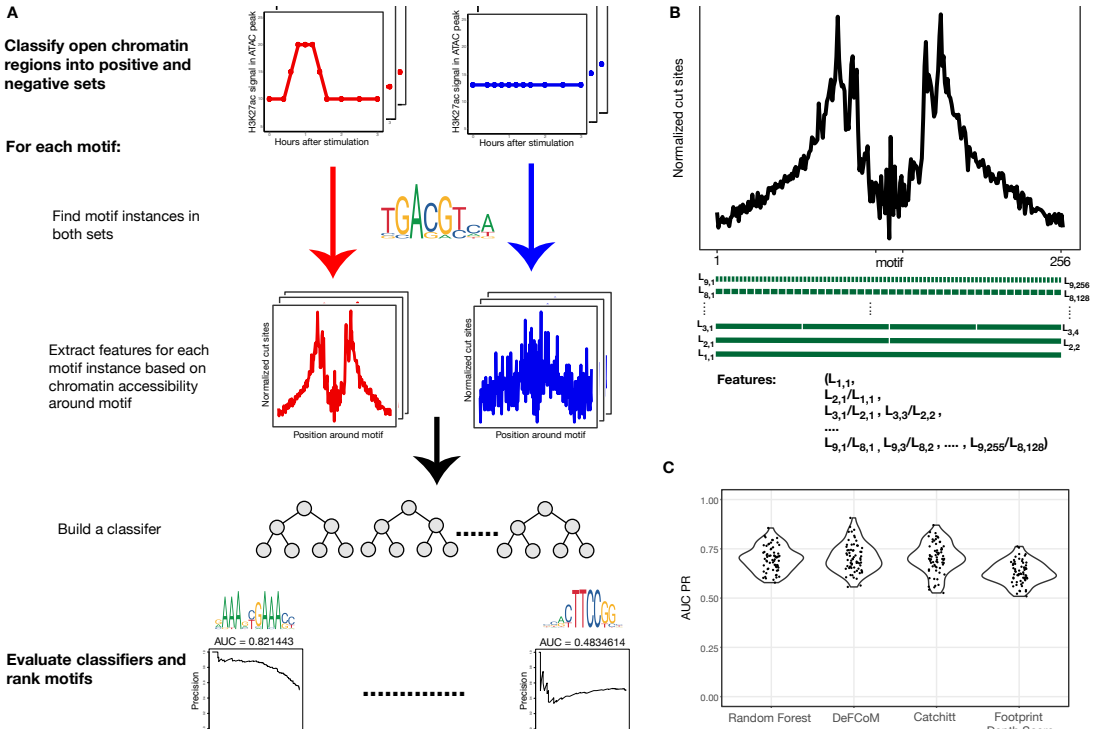


Figure IV-1 | A supervised learning approach to detect functional motifs sequences.

A) Illustration of the pipeline. First, regulatory regions are divided into positive and negative sets based on some features of their activity (e.g. activation time). Then, for each sequence motif, we build a random forest classifier using the cut sites from the chromatin accessibility assay as our features. We then evaluate the success rate for each classifier. High scoring motifs indicate a difference in chromatin structure due to the functionality of the motif in our positive set. B) Illustration of the cut site transformation used as a feature for the classifier. C) Evaluation and comparison of the random forest classifier. AUC PR values of the classifier when used to predict TF binding in 66 TF binding motifs. Results are compared to the AUC PR values of two previously published methods for TF binding prediction - DeFCoM and Catchitt, as well as a simple method of predicting binding with the footprint prediction score. AUC PR values are mean across 5-fold cross-validation

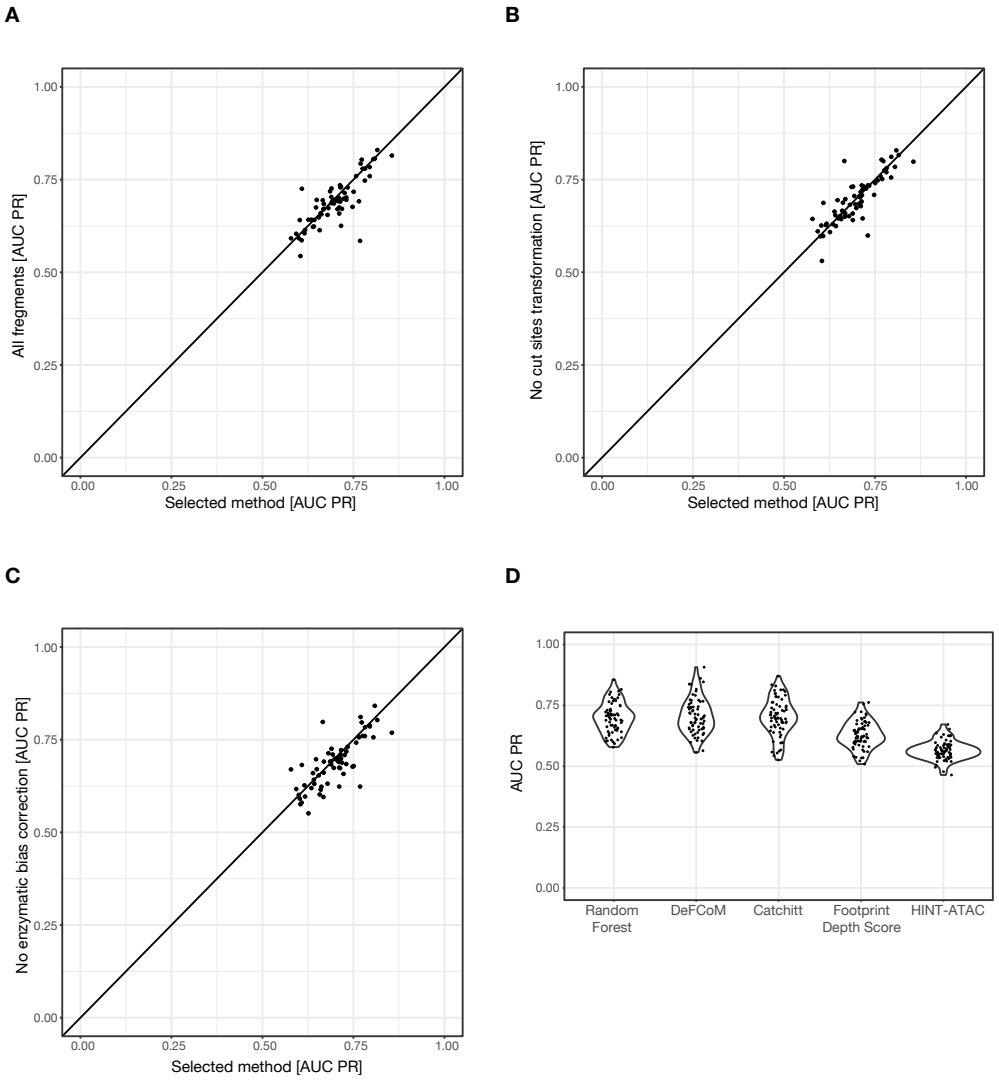


Figure IV-2 | Comparison to existing methods

A) Mean AUC PR values across 5-fold cross-validation of our selected method (x-axis), against the same pipeline but including all read fragments lengths instead of only short fragments (y-axis). B) Similar to A, except the y-axis depicts the same pipeline as our selected method but using the cut site counts in each position as features, instead of the transformed cut sites. C) Similar to A, except the y-axis depicts the same pipeline as our selected method except the input was the cut site count in each position without correcting for enzymatic bias.

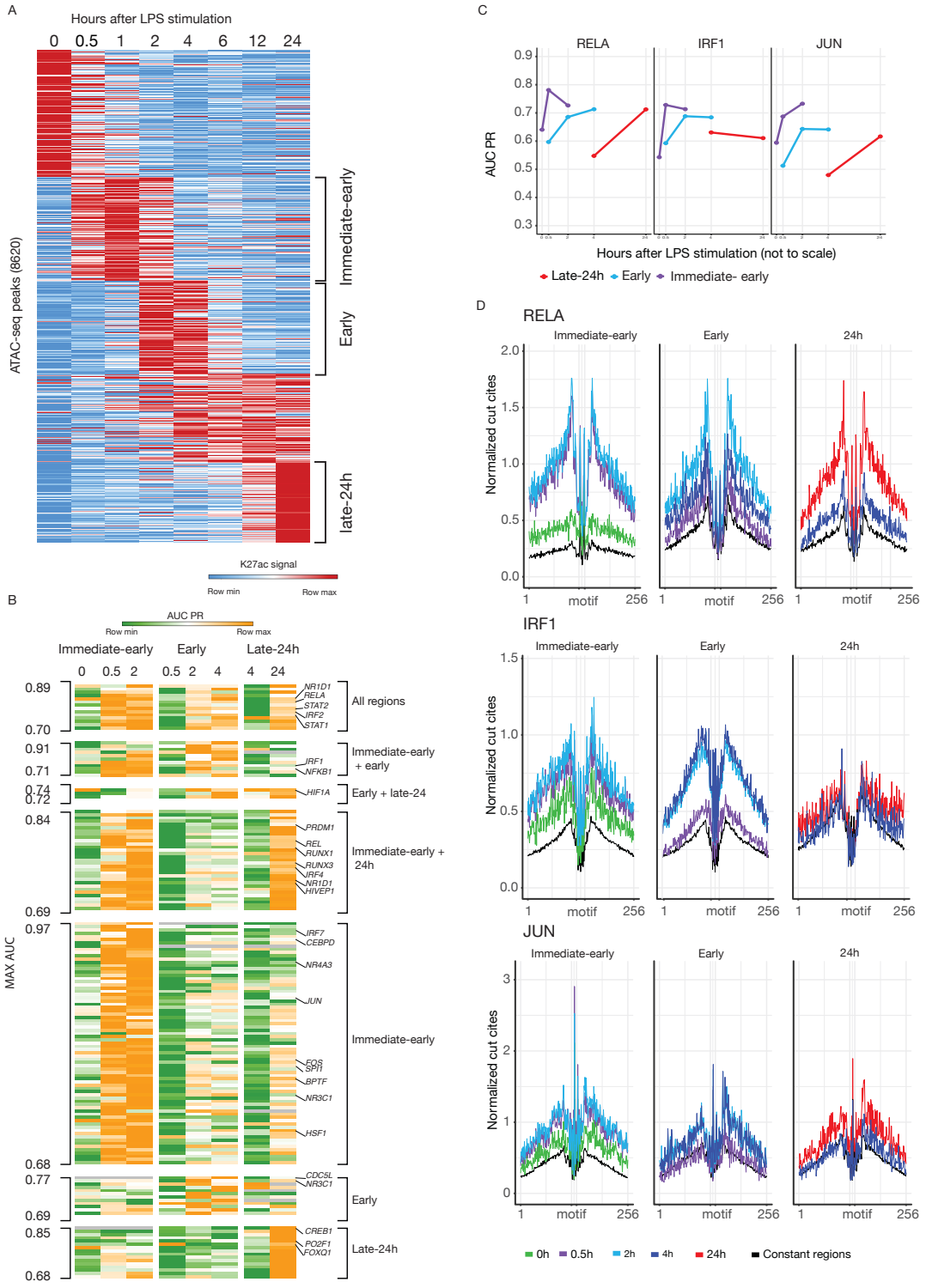


Figure IV-3 | Comprehensive map of predictive TF binding motifs in temporally-

activated regulatory regions.

A) Mean normalized K27 signal across 5 donors in all regulatory regions which exhibit temporal changes in H3K27ac signal across time after LPS stimulation. B) Heatmap summarizing the AUC PR values of all motifs that were found to be significant (FDR < 0.05) in at least one set of regions in one peak time point. Column names are the region's temporal cluster and time after LPS stimulation. Rows are grouped by the set of regions in which the motif was significant, and each group is ordered by the max AUC PR value of each line. C) Plots highlighting the AUC PR values from B for 3 motifs of TFs known to be involved in the DC response to LPS. D) Mean normalized cut sites for each of the motifs from C in all sets of regions before and during peak K27 signal. Cut sites count from the constant regions (black lines) were computed at the earliest time point in each plot.

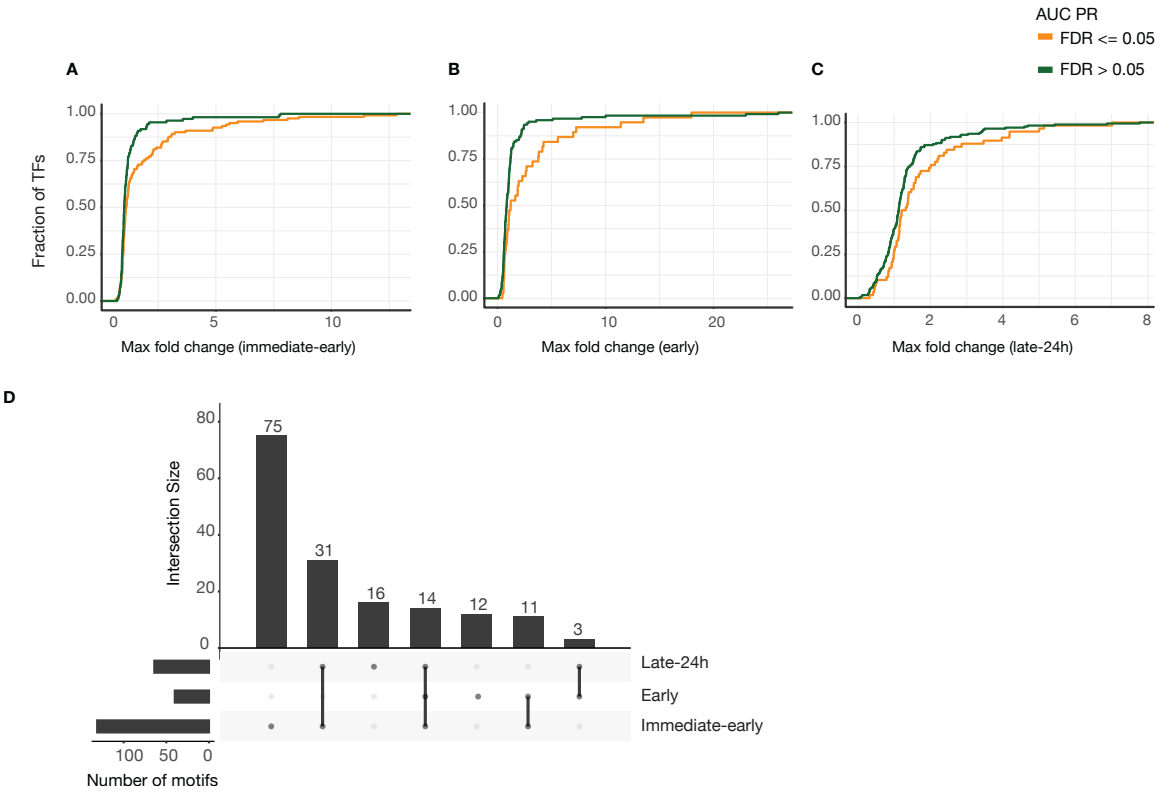


Figure IV-4 | Maximum fold change of TFs that are predictive of H3K27ac signal induction

(A-C) TFs bound in temporally-activated regulatory regions show increase in expression following LPS stimulation. Cumulative distribution function of the maximum TPM fold change from peak activation time points to time point 0 for the immediate-early (A), early (B) and late (C) regions. TFs whose binding motifs were found to be significantly bound (FDR adjusted p-value <= 0.05) in peak activation time points are in orange, while TFs with an FDR adjusted p-value > 0.05 are in green. For time points 30m and 2h, if a motif was significant in one set of regions (e.g. immediate-early) but was not significant at the other set (e.g. early), it was discarded from the analysis for the set of regions in which it was not significant. D) Overlap of TFs that are predictive of H3K27ac signal at different time points

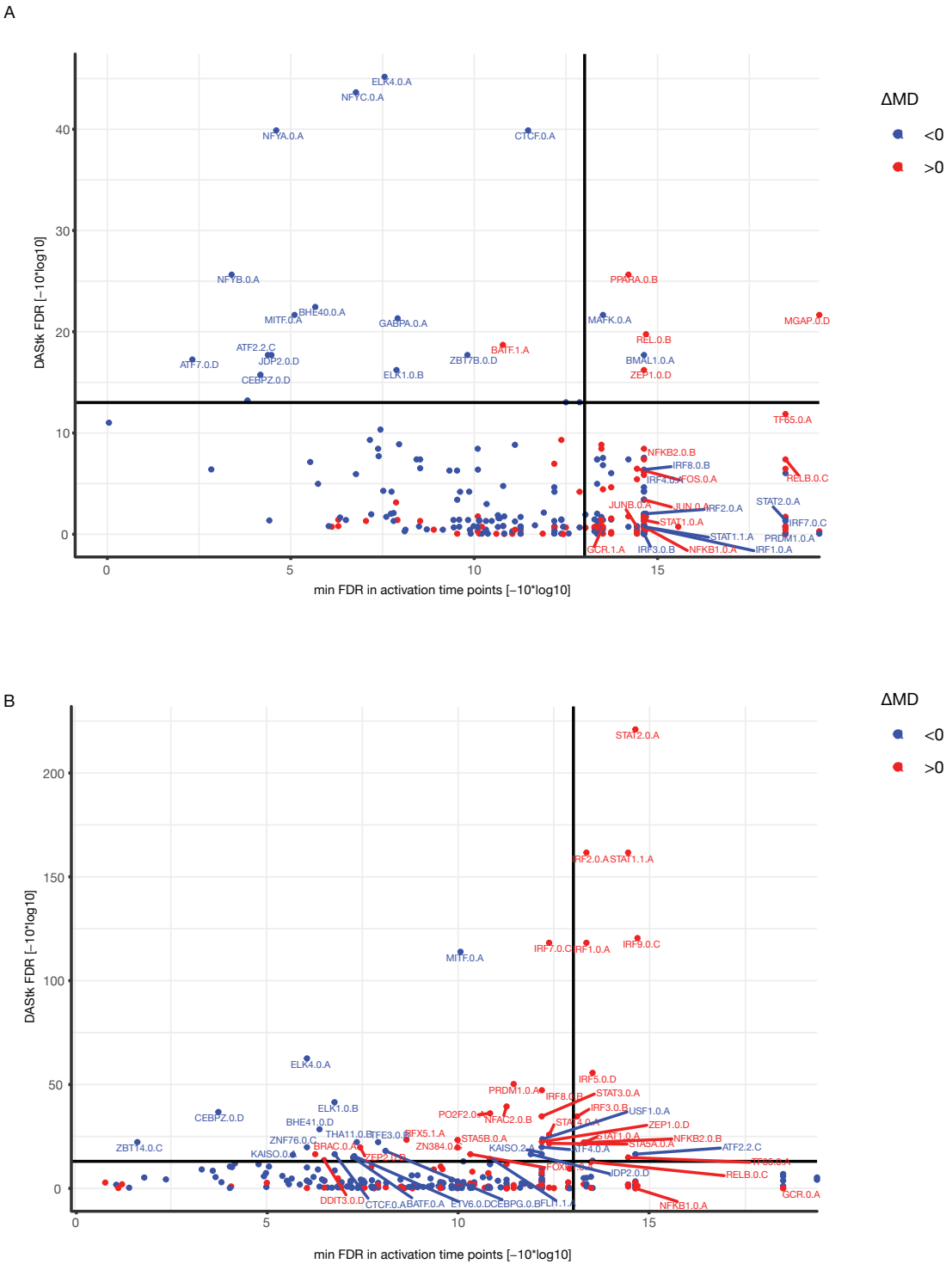


Figure IV-5 | Comparison to DAStk: DAStk results in the set of immediate-

early regions

(A) and early regions (B). Y-axis shows the FDR adjusted p-value of DAStk, while the x-axis shows the minimum of FDR adjusted p-value of our classifier in the activation time points (30m and 2h for the immediate-early regions, 2h and 4h for the early regions). Motifs are colored based on the MD score, where a positive value indicates that this motif is enriched in the induced regions. Horizontal and vertical lines show an FDR value of 0.05. Motifs with an FDR < 0.05 by DAStk and a few other selected motifs are named in the plot.

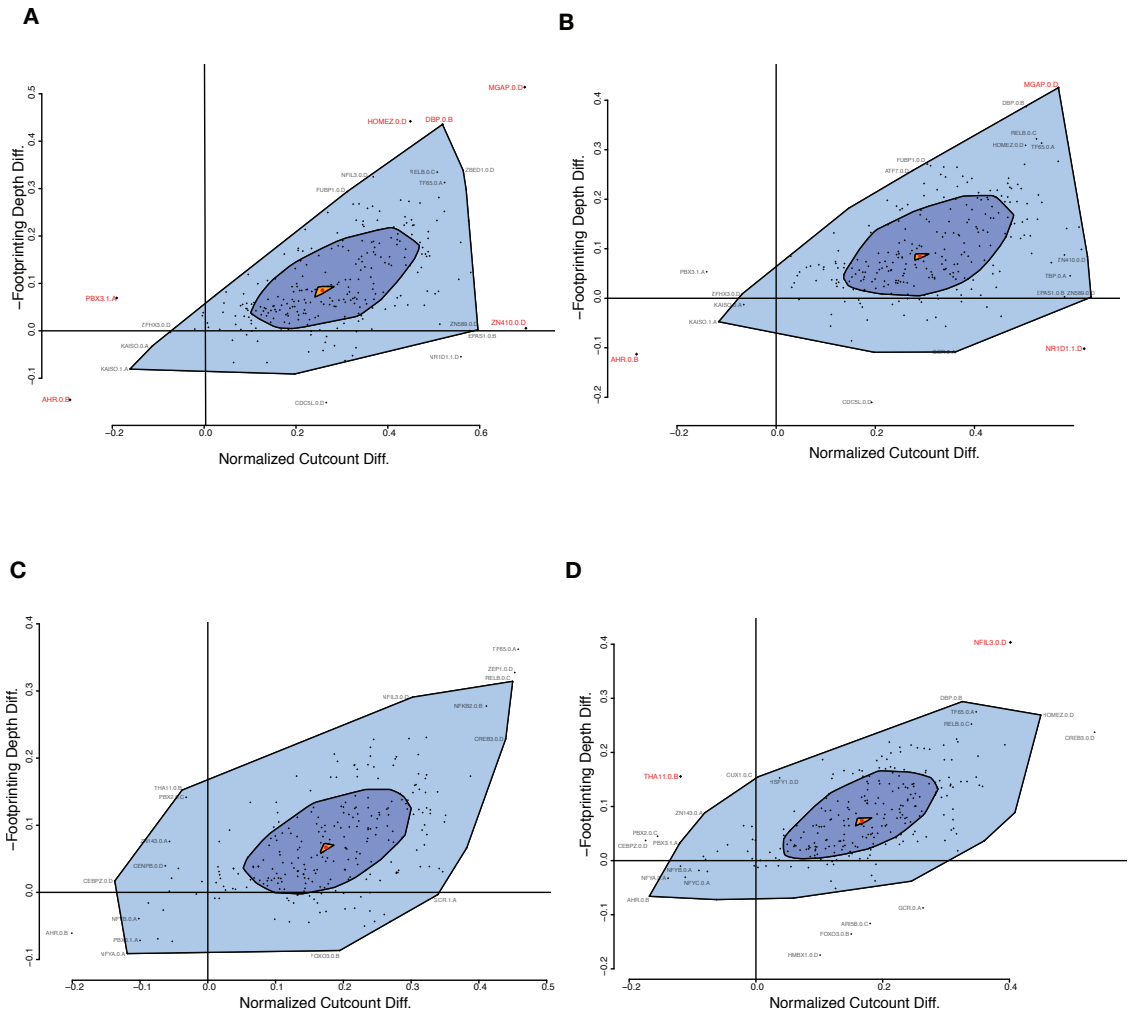


Figure IV-6 | Results of the Bagfoot algorithm

Bag plot depicting the difference in footprint depth and flanking accessibility of each motif in the immediate early regions at 30m (A) and 2h (B), and bag plots for the early regions at 2h (C) and 4h (D). Motifs with a BH corrected p-value <= 0.05 are named in red, while motifs with a BH corrected p-value > 0.05 but a p-value <= 0.05 are named in black.

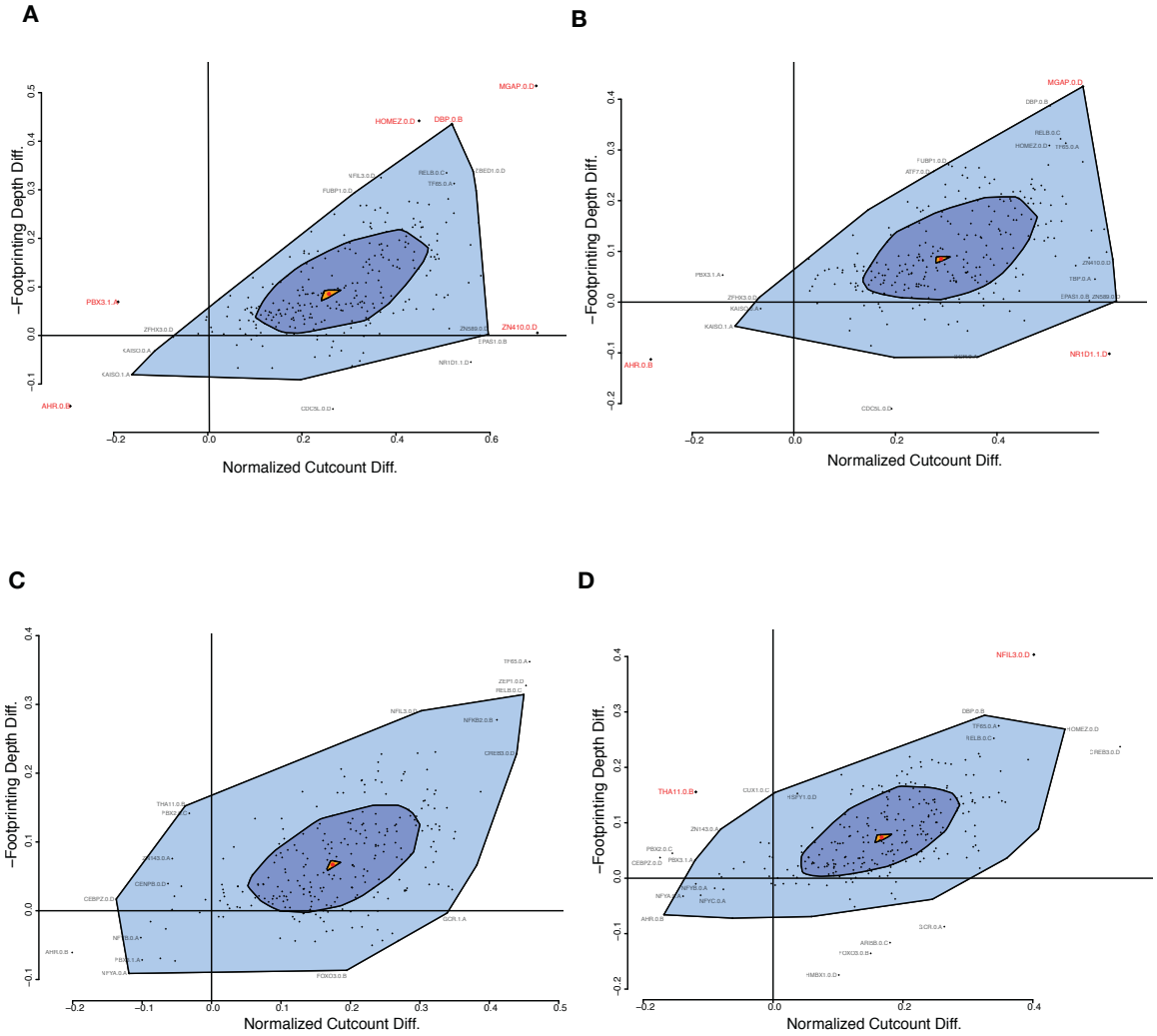


Figure IV-7 | Association between TF binding and H3K27ac signal strength.

A) Enrichment of motifs in immediate-early regions that exhibit a strong change in their H3K27ac signal in donors with a genomic variant. The X-axis shows the p-value of the association test between each motif instance and z-score of the H3K27ac signal (Methods) at three different time points. Y-axis shows the FDR corrected p-value of our classifier at each time point. Motifs with an enrichment p-value < 0.05 and a classification FDR corrected p-value < 0.05 are named in the plot. The size of each point represents the number of immediate-early regions with at least one motif instance. Each point is colored based on the number of regions in which the genetic variants overlap the motif instance. B) Same as A,

except for early induced regions. For the lower plot, in addition to motifs named as in A, we also named a few motifs that only have an enrichment p-value < 0.05 and are known to be associated with histone acetylation.

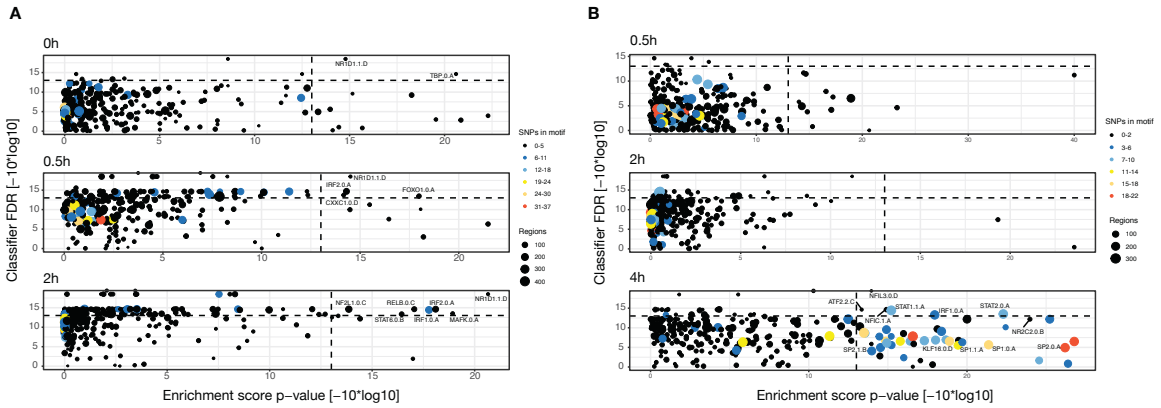
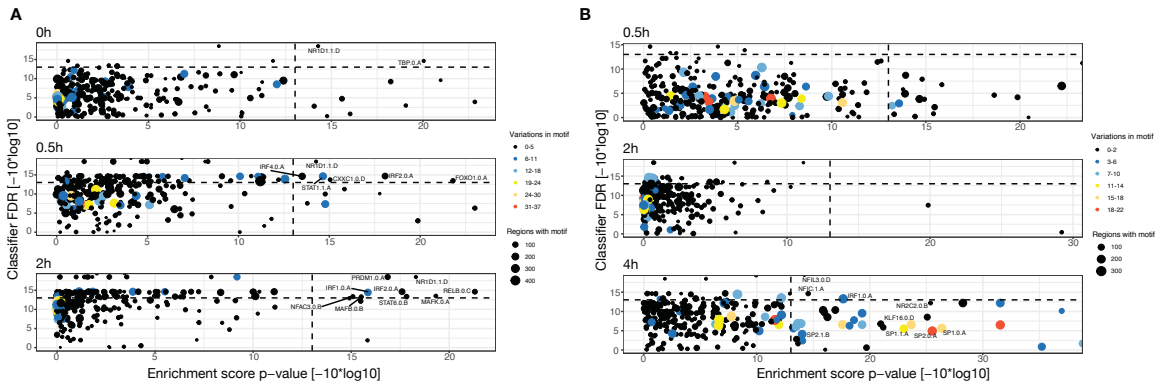


Figure IV-8 | Association between TF binding motifs and H3K27ac signal strength.

A) Enrichment of motifs in immediate-early regions that exhibit a strong change in their H3K27ac signal in donors with a genetic variant. The x-axis shows the p-value of the association test between each motif and the z-scores of the H3K27ac signal (Methods) at three different time points. Y-axis shows the FDR corrected p-value of our classifier at each time point. Motifs with an enrichment p-value < 0.05 and a classification FDR corrected p-value < 0.05 are named in the plot. The size of each point represents the number of immediate-early regions with at least one motif instance. Each point is colored based on the number of regions in which the genetic variants overlap the motif instance. B) Same as A, except for early induced regions. For the 4h plot, in addition to motifs named as in A, we also named a few motifs that only have an enrichment p-value < 0.05 and are known to be associated with histone acetylation and deacetylation.



FigureIV-9 | Association between TF binding and H3K27ac signal strength without the number of motifs in each region as magnitude of increment.

Enrichment of motifs in immediate-early regions (A) and early regions (B) that exhibit a strong change in their H3K27ac signal in donors with a genomic variant. This plot is similar to Figure 4, except that the magnitude of increment is one for each region (instead of the number of motif instances within the region).

V. CHAPTER V: Discussion

V.1. Preface

This chapter is based on my discussions in chapters II, III, and IV

V.2. INTRODUCTION

The first eukaryotic enhancer was identified in 1983 (Mercola et al. 1983) as a 70bp sequence intronic to the immunoglobin heavy chain gene almost 37 years ago, since then we have developed by leaps and bounds creating comprehensive catalogs of putative enhancers elements in many cell types, tissues, and conditions. However, understanding the role of enhancers in regulating gene expression is still a challenge. The major setbacks that we are facing to realize this answer stems from the fact that enhancers are distal elements, about 5 active enhancers for every gene, and associating them to their target gene is not trivial. Over the years many experimental and computational methods were developed to understand what fraction of enhancers are functional and of which how many of them are primary and how many are shadow/ redundant enhancers. However, these methods are limited by the number of loci that can be tested and by the effect size of the individual enhancers. In this thesis, I presented a multi-pronged approach to decipher the role of enhancers in controlling gene expression regulation.

V.3. Comparative Genomics

The results of chapter II demonstrate that most active enhancer elements are not conserved. Our results show that functionally important ones, for example, those that regulate genes with conserved expression, are conserved (Donnard et al. 2018; Danko et al. 2018; Berthelot et al. 2018).

These results relied on a typical phylogeny optimized for maximum branch length, where every pair of species diverged millions of years ago. One of the main challenges of comparative genomics of regulatory elements has been that only a few, highly conserved elements are shared beyond individual species, while the majority of elements are only active in a single node of the tree. In such setup, it is not possible to isolate the impact of an individual enhancer becoming inactive, or the impact of sequence changes on the activity of a given enhancer. Simply, when most elements are node-specific, there is no data to measure the effectiveness of their gain or loss. Further, when most elements that regulate a given gene turn over, it is not possible to isolate the effect of the loss or gain of one specific element on the expression of the associated gene. In short, most prior comparative studies have been grossly underpowered to infer sequence. To overcome these limitations studies focusing on deeply sampling sub-branches are needed.

Another result of chapter II that should be highlighted is the complexity of the regulatory landscape of genes is highly conserved across species and predictive models trained in one species are portable to the other species even with more than 60% of the regulatory elements that are not conserved.

Previously it has been proposed that conserved enhancers might be exerting a stabilizing effect on gene expression levels while the recently evolved enhancers have weak to a neutral role (Kellis et al. 2014; Cooper and Brown 2008). While this hypothesis may be true for a few gene modules, results of chapter II show that even though enhancer elements are not conserved as total units, recently evolved enhancers might be providing the TF binding motifs required for the gene expression, as the model trained in one species is portable to predict gene expression in other. This result is backed up by the fact that using models based on short DNA sequences to predict enhancers in one species perform well in other species (Chen, Fish, and Capra 2018).

V.4. Chromatin Interactions

Results of chapter III reveal that using chromatin interactions I was able to discern the enhancer-promoter interactions with higher accuracy. Even though there have been many computational methods that were developed to identify enriched physical interactions between two genomic regions from the chromatin interaction assays, they have not been able to discern promoter interactions. Comparing the interactions derived from bulk assays and single-cell based assays, it became evident that the interactions are quite variable from cell to cell. This suggests that many interactions will not be represented as strong signals in the contact maps. Microscope-based methods in line with the above showed that enhancer-promoter interactions, in particular, are highly transient in a given pool of cells. And thus it is not surprising that interactions called by existing loop

calling do not detect many enhancer-promoter interactions as they rely heavily on finding discrete punta like regions. Further, the loops identified by these algorithms are heavily enriched in divergent CTCF motifs. Global depletion of CTFC has been shown to have little effect on gene expression misregulation, suggesting that these loops are structural rather than regulatory in nature. The methods developed in chapter III are able to find the local enrichment of signal and thus can be applied to any genomic regions of interest. These interactions proved to have higher accuracy in assigning enhancers to target genes and this resulted in enabling me to build better predictive models of gene expression. However, I will also note that the features used in the model (scores of TF binding motifs per gene) may not fully capture all the regulatory mechanisms of the gene. For example, our model performs very poorly in predicting the expression changes for the majority of downregulated genes. The gene expression for output measured can be due to post-transcriptional regulations. To get a precise readout we need to apply my models to predict the levels of nascent transcripts.

The enhancer-promoter interaction maps, both for consistent and heterogeneous configurations generated in chapter III point us to the potential regulatory elements involved in regulating a given gene. They lack the resolution to explain the extent to which each enhancer is contributing. To get that information we need to devise functional screens for the enhancers and gene of interest. Due to the sheer number of functional elements and the effect size of

the enhancers published, CRISPR based screens will not be suitable. For example, if we want to understand the contribution of each enhancer and all the higher-order interactions for a gene that has 6 enhancers we need to test 6! combinations i.e, 720 possibilities. To test these potential regulatory interactions I propose a new strategy here. We could employ a single cell-based CRISPR screen where we infect cells with a pool of guide RNA viruses that target enhancers of interest for a given gene by controlling molecules of infectivity (MOI) and enriching for the target gene in the single-cell RNA-seq. By modeling the number of viral particles per cell we can determine the optimum number of cells to sequence and the MOI at which the cells should be infected. For example, with MOI of 2 and 1000 cells, we will have enough power to detect each enhancer by itself and all the possible higher-order interactions. Next coming to the effect size of each enhancer or configuration, the regular poly(A) single-cell readout will be too noisy and can only capture the enhancers with high effect size. To overcome this we can use single-cell RNA-seq coupled with target-specific beads. 10x genomics has a commercial kit and there are also a few publications recently that developed target specific beads for in-house methods such as drop-seq or inDrop.

V.5. Remodeling Epigenetic Landscape

Results presented in chapter IV reveal how TF binding is correlated with changes in the activation of putative regulatory elements. The framework we built-in is highly adaptable. The performance of this framework depends on a

good quality motif database, but this can also be extended to any size DNA k-mers. As the resources for TF chip improve we get good motifs that make our frameworks better.

Another possible extension to this framework is to go beyond single motif binding. As we understand now, at cis-regulatory elements many TFs don't bind as monomers but often multiple TFs are bound and these are typically referred to as cis-regulatory modules (CRMs). To this end, we could change the feature matrix on which the classification task is performed to reflect the binding of pairs or Ns of motifs instances within a certain number of base pairs.

And finally, for the TFs we identified as predictive of inducing activity at putative regulatory elements in chapter IV testing the direct effect is imperative. As I mentioned in my introduction it has been shown with a few examples that certain regulatory regions act as billboards while others act as enhanceosomes. The TFs binding in these two situations has largely different outcomes. TFs binding to billboards like regulatory elements, or not binding will not hugely affect their state or transcriptional output, while TFs binding to enhanceosome like regulatory elements will have a dramatic effect. So, the TFs that we prioritized by our framework in Chapter IV should be tested experimentally by knocking out or knocking down the TFs and measuring the genome-wide changes in the H3K27ac signal.

V.6. Conclusion

The work presented in this dissertation provides a framework to understand the role of enhancers in regulating gene expression. This understanding is crucial given that many disease-associated SNPs are found to be in enhancer like regions and understanding how enhancers work will provide a well-positioned understanding of the functional role of various risk variants that are cataloged in genome-wide association studies. Functionally annotating the risk variants in enhancer elements and their target genes is a valuable step for both clinical and basic biology.

VI. Chapter VI: References

Amano, Takanori, Tomoko Sagai, Hideyuki Tanabe, Yoichi Mizushina, Hiromi Nakazawa, and Toshihiko Shiroishi. 2009. “Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription.” Developmental Cell 16 (1): 47–57.

Amano, Takanori, Tomoko Sagai, Hideyuki Tanabe, Yoichi Mizushina, Hiromi Nakazawa, and Toshihiko Shiroishi. 2009. “Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription.” Developmental Cell 16 (1): 47–57.

Ambrosini, Giovanna, Romain Groux, and Philipp Bucher. 2018. “PWMScan: A Fast Tool for Scanning Entire Genomes with a Position-Specific Weight Matrix.” Bioinformatics 34 (14): 2483–84.

Amit, Ido, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, et al. 2009. “Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses.” Science 326 (5950): 257–63.

Amit, Ido, Manuel Garber, Nicolas Chevrier, Ana Paula Leite, Yoni Donner, Thomas Eisenhaure, Mitchell Guttman, et al. 2009. “Unbiased Reconstruction of a Mammalian Transcriptional Network Mediating Pathogen Responses.” Science 326 (5950): 257–63.

Arnosti, David N., and Meghana M. Kulkarni. 2005. “Transcriptional Enhancers: Intelligent Enhanceosomes or Flexible Billboards?” Journal of Cellular Biochemistry 94 (5): 890–98.

Auerbach, Raymond K., Ghia Euskirchen, Joel Rozowsky, Nathan Lamarre-Vincent, Zarmik Moqtaderi, Philippe Lefrançois, Kevin Struhl, Mark Gerstein, and Michael Snyder. 2009. “Mapping Accessible Chromatin Regions Using Sono-Seq.” Proceedings of the National Academy of Sciences of the United States of America 106 (35): 14926–31.

Baek, Songjoon, Ido Goldstein, and Gordon L. Hager. 2017. “Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity.” Cell Reports 19 (8): 1710–22.

Ballester, Benoit, Alejandra Medina-Rivera, Dominic Schmidt, Mar Gonzàlez-Porta, Matthew Carlucci, Xiaoting Chen, Kyle Chessman, et al. 2014. “Multi-Species, Multi-Transcription Factor Binding Highlights Conserved Control of Tissue-Specific Biological Pathways.” eLife 3 (October): e02626.

Bannister, Andrew J., and Tony Kouzarides. 2011. “Regulation of Chromatin by Histone Modifications.” Cell Research 21 (3): 381–95.

Barutcu, A. Rasim, Deli Hong, Bryan R. Lajoie, Rachel Patton McCord, Andre J. van Wijnen, Jane B. Lian, Janet L. Stein, Job Dekker, Anthony N. Imbalzano, and Gary S. Stein. 2016. “RUNX1 Contributes to Higher-Order Chromatin Organization and Gene Regulation in Breast Cancer Cells.” Biochimica et Biophysica Acta 1859 (11): 1389–97.

Beagrie, Robert A., Antonio Scialdone, Markus Schueler, Dorothee C. A. Kraemer, Mita Chotalia, Sheila Q. Xie, Mariano Barbieri, et al. 2017. “Complex Multi-Enhancer Contacts Captured by Genome Architecture Mapping.” Nature. https://doi.org/10.1038/nature21411.

Bejjani, Fabienne, Emilie Evanno, Kazem Zibara, Marc Piechaczyk, and Isabelle Jariel-Encontre. 2019. “The AP-1 Transcriptional Complex: Local Switch or Remote Command?” Biochimica et Biophysica Acta (BBA) - Reviews on Cancer 1872 (1): 11–23.

Beliveau, Brian J., Alistair N. Boettiger, Guy Nir, Bogdan Bintu, Peng Yin, Xiaowei Zhuang, and C-Ting Wu. 2017. “In Situ Super-Resolution Imaging of Genomic DNA with OligoSTORM and OligoDNA-PAINT.” Methods in Molecular Biology 1663: 231–52.

Benabdallah, Nezha S., Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. 2019. “Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation.” Molecular Cell 76 (3): 473–84.e7.

Blashfield, R. K. 1991. “Finding Groups in Data-an Introduction to Cluster-Analysis-Kaufman, L, Rousseeuw, PJ.” SPRINGER VERLAG 175 FIFTH AVE, NEW YORK, NY 10010.

Boekhoudt, Gunther H., Zhu Guo, Guy W. Beresford, and Jeremy M. Boss. 2003. “Communication between NF-Kappa B and Sp1 Controls Histone

Acetylation within the Proximal Promoter of the Monocyte Chemoattractant Protein 1 Gene." Journal of Immunology 170 (8): 4139–47.

Bogdan, C. 2001. "Nitric Oxide and the Immune Response." Nature Immunology 2 (10): 907–16.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." Bioinformatics 30 (15): 2114–20.

Bornstein, Chamutal, Deborah Winter, Zohar Barnett-Itzhaki, Eyal David, Sabah Kadri, Manuel Garber, and Ido Amit. 2014. "A Negative Feedback Loop of Transcription Factors Specifies Alternative Dendritic Cell Chromatin States." Molecular Cell 56 (6): 749–62.

Boukhaled, Giselle M., Mario Corrado, Hannah Guak, and Connie M. Krawczyk. 2019. "Chromatin Architecture as an Essential Determinant of Dendritic Cell Function." Frontiers in Immunology 10 (June): 1119.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." Cell 132 (2): 311–22.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. "ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide." Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.] 109 (January): 21.29.1–9.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” Nature Methods 10 (12): 1213–18.

Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. “Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position.” Nature Methods, October. https://doi.org/10.1038/nmeth.2688.

Calo, Eliezer, and Joanna Wysocka. 2013. “Modification of Enhancer Chromatin: What, How, and Why?” Molecular Cell 49 (5): 825–37.

Cannavò, Enrico, Pierre Khoueiry, David A. Garfield, Paul Geeleher, Thomas Zichner, E. Hilary Gustafson, Lucia Ciglar, Jan O. Korbel, and Eileen E. M. Furlong. 2016. “Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks.” Current Biology: CB 26 (1): 38–51.

Chavanas, Stéphane, Véronique Adoue, Marie-Claire Méchin, Shibo Ying, Sijun Dong, Hélène Duplan, Marie Charveron, Hidenari Takahara, Guy Serre, and Michel Simon. 2008. “Long-Range Enhancer Associated with Chromatin Looping Allows AP-1 Regulation of the Peptidylarginine Deiminase 3 Gene in Differentiated Keratinocyte.” PloS One 3 (10): e3408.

Chen, Jenny, Alexander A. Shishkin, Xiaopeng Zhu, Sabah Kadri, Itay Maza, Mitchell Guttman, Jacob H. Hanna, Aviv Regev, and Manuel Garber. 2016.

“Evolutionary Analysis across Mammals Reveals Distinct Classes of Long Non-Coding RNAs.” Genome Biology 17 (1): 19.

Cheng, Yong, Weisheng Wu, Swathi Ashok Kumar, Duonan Yu, Wulan Deng, Tamara Tripic, David C. King, et al. 2009. “Erythroid GATA1 Function Revealed by Genome-Wide Analysis of Transcription Factor Occupancy, Histone Modifications, and mRNA Expression.” Genome Research 19 (12): 2172–84.

Cheng, Yong, Zhihai Ma, Bong-Hyun Kim, Weisheng Wu, Philip Cayting, Alan P. Boyle, Vasavi Sundaram, et al. 2014. “Principles of Regulatory Information Conservation between Mouse and Human.” Nature 515 (7527): 371–75.

Chew, Joon-Lin, Yuin-Han Loh, Wensheng Zhang, Xi Chen, Wai-Leong Tam, Leng-Siew Yeap, Pin Li, et al. 2005. “Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells.” Molecular and Cellular Biology 25 (14): 6031–46.

Chinenov, Yurii, Maddalena Coppo, Rebecca Gupte, Maria A. Sacta, and Inez Rogatsky. 2014. “Glucocorticoid Receptor Coordinates Transcription Factor-Dominated Regulatory Network in Macrophages.” BMC Genomics 15 (August): 656.

Chinenov, Yurii, Rebecca Gupte, and Inez Rogatsky. 2013. “Nuclear Receptors in Inflammation Control: Repression by GR and beyond.” Molecular and Cellular Endocrinology 380 (1-2): 55–64.

Claussnitzer, Melina, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, et al. 2015. “FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.” The New England Journal of Medicine 373 (10): 895–907.

Cramer, Thorsten, Yuji Yamanishi, Björn E. Clausen, Irmgard Förster, Rafal Pawlinski, Nigel Mackman, Volker H. Haase, et al. 2003. “HIF-1alpha Is Essential for Myeloid Cell-Mediated Inflammation.” Cell 112 (5): 645–57.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. “Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State.” Proceedings of the National Academy of Sciences of the United States of America 107 (50): 21931–36.

Crocker, Justin, and Albert Erives. 2008. “A Closer Look at the Eve Stripe 2 Enhancers of Drosophila and Themira.” PLoS Genetics 4 (11): e1000276.

Cui, Shuaiying, Katarzyna E. Kolodziej, Naoshi Obara, Alexandra Amaral-Psarris, Jeroen Demmers, Lihong Shi, James Douglas Engel, Frank Grosveld, John Strouboulis, and Osamu Tanabe. 2011. “Nuclear Receptors TR2 and TR4 Recruit Multiple Epigenetic Transcriptional Corepressors That Associate Specifically with the Embryonic β-Type Globin Promoters in Differentiated Adult Erythroid Cells.” Molecular and Cellular Biology 31 (16): 3298–3311.

Daftary, Gaurang S., Gwen A. Lomberk, Navtej S. Buttar, Thomas W. Allen, Adrienne Grzenda, Jinsan Zhang, Ye Zheng, et al. 2012. “Detailed

Structural-Functional Analysis of the Krüppel-like Factor 16 (KLF16) Transcription Factor Reveals Novel Mechanisms for Silencing Sp/KLF Sites Involved in Metabolism and Endocrinology." The Journal of Biological Chemistry 287 (10): 7010–25.

Danko, Charles G., Lauren A. Choate, Brooke A. Marks, Edward J. Rice, Zhong Wang, Tinyi Chu, Andre L. Martins, et al. 2018. "Dynamic Evolution of Regulatory Element Ensembles in Primate CD4+ T Cells." Nature Ecology & Evolution 2 (3): 537–48.

Davis, Carrie A., Benjamin C. Hitz, Cricket A. Sloan, Esther T. Chan, Jean M. Davidson, Idan Gabdank, Jason A. Hilton, et al. 2018. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update." Nucleic Acids Research 46 (D1): D794–801.

Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002. "Capturing Chromosome Conformation." Science 295 (5558): 1306–11.

Dekker, Job, Marc A. Marti-Renom, and Leonid A. Mirny. 2013. "Exploring the Three-Dimensional Organization of Genomes: Interpreting Chromatin Interaction Data." Nature Reviews. Genetics 14 (6): 390–403.

Dekker, Job. 2016. "Mapping the 3D Genome: Aiming for Consilience." Nature Reviews. Molecular Cell Biology 17 (12): 741–42.

Derr, Alan, Chaoxing Yang, Rapolas Zilionis, Alexey Sergushichev, David M. Blodgett, Sambra Redick, Rita Bortell, et al. 2016. "End Sequence Analysis

ToolKit (ESAT) Expands the Extractable from Single Cell RNA-Seq Experiments." Genome Research, July. https://doi.org/10.1101/gr.207902.116.

Dickel, Diane E., Yiwen Zhu, Alex S. Nord, John N. Wylie, Jennifer A. Akiyama, Veena Afzal, Ingrid Plajzer-Frick, et al. 2014. "Function-Based Identification of Mammalian Enhancers Using Site-Specific Integration." Nature Methods, March. https://doi.org/10.1038/nmeth.2886.

Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." Nature 485 (7398): 376–80.

Doetzlhofer, A., H. Rotheneder, G. Lagger, M. Koranda, V. Kurtev, G. Brosch, E. Wintersberger, and C. Seiser. 1999. "Histone Deacetylase 1 Can Repress Transcription by Binding to Sp1." Molecular and Cellular Biology 19 (8): 5504–11.

Donnard, Elisa, Pranitha Vangala, Shaked Afik, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, et al. 2018. "Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon." Cell Systems 6 (3): 381–94.e7.

Donnard, Elisa, Pranitha Vangala, Shaked Afik, Sean McCauley, Anetta Nowosielska, Alper Kucukural, Barbara Tabak, et al. 2018. "Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon." Cell Systems, January. https://doi.org/10.1016/j.cels.2018.01.002.

Dostie, Josée, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, et al. 2006. “Chromosome Conformation Capture Carbon Copy (5C): A Massively Parallel Solution for Mapping Interactions between Genomic Elements.” Genome Research 16 (10): 1299–1309.

Dunipace, Leslie, Anil Ozdemir, and Angelike Stathopoulos. 2011. “Complex Interactions between Cis-Regulatory Modules in Native Conformation Are Critical for Drosophila Snail Expression.” Development 138 (18): 4075–84.

Edelman, Lucas Brandon, and Peter Fraser. 2012. “Transcription Factories: Genetic Programming in Three Dimensions.” Current Opinion in Genetics & Development 22 (2): 110–14.

El Khattabi, Laila, Haiyan Zhao, Jens Kalchschmidt, Natalie Young, Seolkyoung Jung, Peter Van Blerkom, Philippe Kieffer-Kwon, et al. 2019. “A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers.” Cell 178 (5): 1145–58.e20.

ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” Nature 489 (7414): 57–74.

Escoubet-Lozach, Laure, Christopher Benner, Minna U. Kaikkonen, Jean Lozach, Sven Heinz, Nathan J. Spann, Andrea Crotti, et al. 2011. “Mechanisms Establishing TLR4-Responsive Activation States of Inflammatory Response Genes.” PLoS Genetics 7 (12): e1002401.

Fan, Wuqiang, Hidetaka Morinaga, Jane J. Kim, Eunju Bae, Nathanael J. Spann, Sven Heinz, Christopher K. Glass, and Jerrold M. Olefsky. 2010. “FoxO1 Regulates Tlr4 Inflammatory Pathway Signalling in Macrophages.” The EMBO Journal 29 (24): 4223–36.

Fischer, David S., Fabian J. Theis, and Nir Yosef. 2018. “Impulse Model-Based Differential Expression Analysis of Time Course Sequencing Data.” Nucleic Acids Research 46 (20): e119.

Fontaine, Coralie, Elena Rigamonti, Benoit Pourcet, Hélène Duez, Christian Duhem, Jean-Charles Fruchart, Giulia Chinetti-Gbaguidi, and Bart Staels. 2008. “The Nuclear Receptor Rev-Erbalpha Is a Liver X Receptor (LXR) Target Gene Driving a Negative Feedback Loop on Select LXR-Induced Pathways in Human Macrophages.” Molecular Endocrinology 22 (8): 1797–1811.

Fowler, Trent, Ranjan Sen, and Ananda L. Roy. 2011. “Regulation of Primary Response Genes.” Molecular Cell 44 (3): 348–60.

Frey, Wesley D., Anisha Chaudhry, Priscila F. Slepicka, Adam M. Ouellette, Steven E. Kirberger, William C. K. Pomerantz, Gregory J. Hannon, and Camila O. Dos Santos. 2017. “BPTF Maintains Chromatin Accessibility and the Self-Renewal Capacity of Mammary Gland Stem Cells.” Stem Cell Reports 9 (1): 23–31.

Fukaya, Takashi, Bomyi Lim, and Michael Levine. 2016. “Enhancer Control of Transcriptional Bursting.” Cell 166 (2): 358–68.

Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. n.d. “Activity-by-Contact Model of Enhancer Specificity from Thousands of CRISPR Perturbations.” https://doi.org/10.1101/529990.

Furlong, Eileen E. M., and Michael Levine. 2018. “Developmental Enhancers and Chromosome Topology.” Science 361 (6409): 1341–45.

Gallagher, Michael D., and Alice S. Chen-Plotkin. 2018. “The Post-GWAS Era: From Association to Function.” American Journal of Human Genetics 102 (5): 717–30.

Garber, Manuel, Mitchell Guttman, Michele Clamp, Michael C. Zody, Nir Friedman, and Xiaohui Xie. 2009. “Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns.” Bioinformatics 25 (12): i54–62.

Garber, Manuel, Nir Yosef, Alon Goren, Raktima Raychowdhury, Anne Thielke, Mitchell Guttman, James Robinson, et al. 2012. “A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals.” Molecular Cell 47 (5): 810–22.

Gasperini, Molly, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, et al. 2019. “A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.” Cell 176 (6): 1516.

Gerritsen, M. E., A. J. Williams, A. S. Neish, S. Moore, Y. Shi, and T. Collins. 1997. “CREB-Binding protein/p300 Are Transcriptional Coactivators of

p65.” Proceedings of the National Academy of Sciences of the United States of America 94 (7): 2927–32.

Ghandi, Mahmoud, Morteza Mohammad-Noori, Narges Ghareghani, Dongwon Lee, Levi Garraway, and Michael A. Beer. 2016. “gkmSVM: An R Package for Gapped-Kmer SVM.” Bioinformatics 32 (14): 2205–7.

Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. “Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression.” Nature Genetics 51 (8): 1272–82.

Ghavi-Helm, Yad, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E. M. Furlong. 2014. “Enhancer Loops Appear Stable during Development and Are Associated with Paused Polymerase.” Nature 512 (7512): 96–100.

Gibbs, Julie E., John Blaikley, Stephen Beesley, Laura Matthews, Karen D. Simpson, Susan H. Boyce, Stuart N. Farrow, et al. 2012. “The Nuclear Receptor REV-ERBα Mediates Circadian Regulation of Innate Immunity through Selective Regulation of Inflammatory Cytokines.” Proceedings of the National Academy of Sciences of the United States of America 109 (2): 582–87.

Gilad, Yoav, Alicia Oshlack, and Scott A. Rifkin. 2006. “Natural Selection on Gene Expression.” Trends in Genetics: TIG 22 (8): 456–61.

Giorgetti, Luca, and Edith Heard. 2016. “Closing the Loop: 3C versus DNA FISH.” Genome Biology 17 (1): 215.

Gong, Q., and A. Dean. 1993. “Enhancer-Dependent Transcription of the Epsilon-Globin Promoter Requires Promoter-Bound GATA-1 and Enhancer-Bound AP-1/NF-E2.” Molecular and Cellular Biology 13 (2): 911–17.

González, Alvaro J., Manu Setty, and Christina S. Leslie. 2015. “Early Enhancer Establishment and Regulatory Locus Complexity Shape Transcriptional Programs in Hematopoietic Differentiation.” Nature Genetics 47 (11): 1249–59.

González, Alvaro J., Manu Setty, and Christina S. Leslie. 2015. “Early Enhancer Establishment and Regulatory Locus Complexity Shape Transcriptional Programs in Hematopoietic Differentiation.” Nature Publishing Group 47 (11): 1249–59.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” Bioinformatics 27 (7): 1017–18.

Grau, Jan, Ivo Grosse, and Jens Keilwagen. 2015. “PRROC: Computing and Visualizing Precision-Recall and Receiver Operating Characteristic Curves in R.” Bioinformatics 31 (15): 2595–97.

Gupta, Shobhit, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. 2007. “Quantifying Similarity between Motifs.” Genome Biology 8 (2): R24.

Gusmao, Eduardo G., Manuel Allhoff, Martin Zenke, and Ivan G. Costa. 2016. “Analysis of Computational Footprinting Methods for DNase Sequencing Experiments.” Nature Methods 13 (4): 303–9.

Guttman, Mitchell, Manuel Garber, Joshua Z. Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, et al. 2010. “Ab Initio Reconstruction of Cell Type–specific Transcriptomes in Mouse Reveals the Conserved Multi-Exonic Structure of lincRNAs.” Nature Biotechnology 28 (5): 503–10.

Harding, Heather P., Yuhong Zhang, Huiquing Zeng, Isabel Novoa, Phoebe D. Lu, Marcella Calfon, Navid Sadri, et al. 2003. “An Integrated Stress Response Regulates Amino Acid Metabolism and Resistance to Oxidative Stress.” Molecular Cell 11 (3): 619–33.

He, Qiye, Anaïs F. Bardet, Brianne Patton, Jennifer Purvis, Jeff Johnston, Ariel Paulson, Madelaine Gogol, Alexander Stark, and Julia Zeitlinger. 2011. “High Conservation of Transcription Factor Binding and Evidence for Combinatorial Regulation across Six Drosophila Species.” Nature Genetics 43 (5): 414–20.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. “Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression.” Nature 459 (7243): 108–12.

Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. “Distinct and

Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome.” Nature Genetics 39 (3): 311–18.

Helft, Julie, Jan Böttcher, Probir Chakravarty, Santiago Zelenay, Jatta Huotari, Barbara U. Schraml, Delphine Goubau, and Caetano Reis e Sousa. 2015. “GM-CSF Mouse Bone Marrow Cultures Comprise a Heterogeneous Population of CD11c+MHCII+ Macrophages and Dendritic Cells.” Immunity 42 (6): 1197–1211.

Hewish, D. R., and L. A. Burgoyne. 1973. “Chromatin Sub-Structure. The Digestion of Chromatin DNA at Regularly Spaced Sites by a Nuclear Deoxyribonuclease.” Biochemical and Biophysical Research Communications 52 (2): 504–10.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, et al. 2006. “The UCSC Genome Browser Database: Update 2006.” Nucleic Acids Research 34 (Database issue): D590–98.

Hoeijmakers, Wieteke Anna Maria, and Richárd Bártfai. 2018. “Characterization of the Nucleosome Landscape by Micrococcal Nuclease-Sequencing (MNase-Seq).” Methods in Molecular Biology 1689: 83–101.

Holde, Kensal E. van. 2012. Chromatin. Springer Science & Business Media.

Hoogenkamp, Maarten, Monika Lichtinger, Hanna Krysinska, Christophe Lancrin, Deborah Clarke, Andrew Williamson, Luca Mazzarella, et al. 2009. “Early Chromatin Unfolding by RUNX1: A Molecular Explanation for Differential

Requirements during Specification versus Maintenance of the Hematopoietic Gene Expression Program." Blood. https://doi.org/10.1182/blood-2008-11-191890.

Hui, Zhaoyuan, Lina Zhou, Zhonghui Xue, Lingfeng Zhou, Yikai Luo, Feng Lin, Xia Liu, et al. 2018. "Cxxc Finger Protein 1 Positively Regulates GM-CSF-Derived Macrophage Phagocytosis Through Csf2rα-Mediated Signaling." Frontiers in Immunology 9 (August): 1885.

Hwang, Yu-Jin, Eun-Woo Lee, Jaewhan Song, Haeng-Ran Kim, Young-Chun Jun, and Kyung-A Hwang. 2013. "MafK Positively Regulates NF-κB Activity by Enhancing CBP-Mediated p65 Acetylation." Scientific Reports 3 (November): 3242.

Icardi, Laura, Raffaele Mori, Viola Gesellchen, Sven Eyckerman, Lode De Cauwer, Judith Verhelst, Koen Vercauteren, et al. 2012. "The Sin3a Repressor Complex Is a Master Regulator of STAT Transcriptional Activity." Proceedings of the National Academy of Sciences of the United States of America 109 (30): 12058–63.

Jeng, Mark Y., Maxwell R. Mumbach, Jeffrey M. Granja, Ansuman T. Satpathy, Howard Y. Chang, and Anne Lynn S. Chang. 2019. "Enhancer Connectome Nominates Target Genes of Inherited Risk Variants from Inflammatory Skin Disorders." The Journal of Investigative Dermatology 139 (3): 605–14.

Ji, Hongkai, Hui Jiang, Wenxiu Ma, David S. Johnson, Richard M. Myers, and Wing H. Wong. 2008. “An Integrated Software System for Analyzing ChIP-Chip and ChIP-Seq Data.” Nature Biotechnology 26 (11): 1293–1300.

Jin, Fulai, Yan Li, Jesse R. Dixon, Siddarth Selvaraj, Zhen Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza, and Bing Ren. 2013. “A High-Resolution Map of the Three-Dimensional Chromatin Interactome in Human Cells.” Nature 503 (7475): 290–94.

Jjingo, Daudi, Andrew B. Conley, Jianrong Wang, Leonardo Mariño-Ramírez, Victoria V. Lunyak, and I. King Jordan. 2014. “Mammalian-Wide Interspersed Repeat (MIR)-Derived Enhancers and the Regulation of Human Gene Expression.” Mobile DNA 5 (May): 14.

John, Sam, Peter J. Sabo, Robert E. Thurman, Myong-Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos. 2011. “Chromatin Accessibility Pre-Determines Glucocorticoid Receptor Binding Patterns.” Nature Genetics 43 (3): 264–68.

Johnson, Jarrod S., Nicholas De Veaux, Alexander W. Rives, Xavier Lahaye, Sasha Y. Lucas, Brieuc P. Perot, Marine Luka, et al. 2020. “A Comprehensive Map of the Monocyte-Derived Dendritic Cell Transcriptional Network Engaged upon Innate Sensing of HIV.” Cell Reports 30 (3): 914–31.e9.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic. 2007. “Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods.” Biostatistics 8 (1): 118–27.

Junion, Guillaume, Mikhail Spivakov, Charles Girardot, Martina Braun, E. Hilary Gustafson, Ewan Birney, and Eileen E. M. Furlong. 2012. “A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History.” Cell 148 (3): 473–86.

Juric, Ivan, Miao Yu, Armen Abnousi, Ramya Raviram, Rongxin Fang, Yuan Zhao, Yanxiao Zhang, et al. 2019. “MAPS: Model-Based Analysis of Long-Range Chromatin Interactions from PLAC-Seq and HiChIP Experiments.” PLoS Computational Biology 15 (4): e1006982.

Kagey, Michael H., Jamie J. Newman, Steve Bilodeau, Ye Zhan, David A. Orlando, Nynke L. van Berkum, Christopher C. Ebmeier, et al. 2010. “Mediator and Cohesin Connect Gene Expression and Chromatin Architecture.” Nature 467 (7314): 430–35.

Katto, Judith, Nicole Engel, Wasim Abbas, Georges Herbein, and Ulrich Mahlknecht. 2013. “Transcription Factor NFκB Regulates the Expression of the Histone Deacetylase SIRT1.” Clinical Epigenetics 5 (1): 11.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. “Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation.” Nature 477 (7364): 289–94.

Keilwagen, Jens, Stefan Posch, and Jan Grau. 2019. “Accurate Prediction of Cell Type-Specific Transcription Factor Binding.” Genome Biology 20 (1): 9.

Keniry, Megan, Maira M. Pires, Sarah Mense, Celine Lefebvre, Boyi Gan, Karen Justiano, Ying-Ka Ingar Lau, et al. 2013. “Survival Factor NFIL3 Restricts

FOXO-Induced Gene Expression in Cancer.” Genes & Development 27 (8): 916–27.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L. Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” Genome Biology 14 (4): R36.

Kim, Min Young, Ji Eun Lee, Lark Kyun Kim, and Taesoo Kim. 2019. “Epigenetic Memory in Gene Regulation and Immune Response.” BMB Reports 52 (2): 127–32.

Ko, Chiung-Yuan, Wen-Chang Chang, and Ju-Ming Wang. 2015. “Biological Roles of CCAAT/Enhancer-Binding Protein Delta during Inflammation.” Journal of Biomedical Science 22 (January): 6.

Koch, Christoph M., Robert M. Andrews, Paul Flicek, Shane C. Dillon, Ulaş Karaöz, Gayle K. Clelland, Sarah Wilcox, et al. 2007. “The Landscape of Histone Modifications across 1% of the Human Genome in Five Human Cell Lines.” Genome Research 17 (6): 691–707.

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, and Others. n.d. “Caret: Classification and Regression Training, 2011.” R Package Version 4.

Kuhn, Max. 2008. “Building Predictive Models in R Using the Caret Package.” Journal of Statistical Software, Articles 28 (5): 1–26.

Kulakovskiy, Ivan V., Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, et al. 2018. “HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis.” Nucleic Acids Research 46 (D1): D252–59.

Kunarso, Galih, Na-Yu Chia, Justin Jeyakani, Catalina Hwang, Xinyi Lu, Yun-Shen Chan, Huck-Hui Ng, and Guillaume Bourque. 2010. “Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells.” Nature Genetics 42 (7): 631–34.

Kutter, Claudia, Stephen Watt, Klara Stefflova, Michael D. Wilson, Angela Goncalves, Chris P. Ponting, Duncan T. Odom, and Ana C. Marques. 2012. “Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression.” PLoS Genetics 8 (7): e1002841.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods 9 (4): 357–59.

Lee, Jeong-Heon, and David G. Skalnik. 2005. “CpG-Binding Protein (CXXC Finger Protein 1) Is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex.” The Journal of Biological Chemistry 280 (50): 41725–31.

Leek, Jeffrey T., W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey. 2012. “The Sva Package for Removing Batch Effects and Other

Unwanted Variation in High-Throughput Experiments." Bioinformatics 28 (6): 882–83.

Lettice, Laura A., Simon J. H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, Ben A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. 2003. "A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly." Human Molecular Genetics 12 (14): 1725–35.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." BMC Bioinformatics 12 (August): 323.

Li, Guoliang, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, et al. 2012. "Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation." Cell 148 (1-2): 84–98.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." Bioinformatics 25 (16): 2078–79.

Li, X., and M. Noll. 1994. "Compatibility between Enhancers and Promoters Determines the Transcriptional Specificity of Gooseberry and Gooseberry Neuro in the Drosophila Embryo." The EMBO Journal 13 (2): 400–406.

Li, Yan, Chloe M. Rivera, Haruhiko Ishii, Fulai Jin, Siddarth Selvaraj, Ah Young Lee, Jesse R. Dixon, and Bing Ren. 2014. “CRISPR Reveals a Distal Super-Enhancer Required for Sox2 Expression in Mouse Embryonic Stem Cells.” PloS One 9 (12): e114485.

Li, Zhijian, Marcel H. Schulz, Thomas Look, Matthias Begemann, Martin Zenke, and Ivan G. Costa. 2019. “Identification of Transcription Factor Binding Sites Using ATAC-Seq.” Genome Biology 20 (1): 45.

Liaw, Andy, Matthew Wiener, and Others. 2002. “Classification and Regression by randomForest.” R News 2 (3): 18–22.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” Science 326 (5950): 289–93.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, et al. 2011. “A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals.” Nature 478 (7370): 476–82.

Link, Verena M., Sascha H. Duttke, Hyun B. Chun, Inge R. Holtman, Emma Westin, Marten A. Hoeksema, Yohei Abe, et al. 2018. “Analysis of Genetically Diverse Macrophages Reveals Local and Domain-Wide Mechanisms That Control Transcription Factor Binding and Function.” Cell 173 (7): 1796–1809.e17.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology 15 (12): 550.

Love, Michael, Simon Anders, and Wolfgang Huber. 2014. “Differential Analysis of Count Data--the DESeq2 Package.” Genome Biology 15: 550.

Lowe, Craig B., and David Haussler. 2012. “29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome.” PloS One 7 (8): e43128.

Ludwig, Leif S., Caleb A. Lareau, Erik L. Bao, Satish K. Nandakumar, Christoph Muus, Jacob C. Ulirsch, Kaitavjeet Chowdhary, et al. 2019. “Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis.” Cell Reports 27 (11): 3228–40.e7.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. “Evidence for Stabilizing Selection in a Eukaryotic Enhancer Element.” Nature 403 (6769): 564–67.

Luxburg, Ulrike von. 2007. “A Tutorial on Spectral Clustering.” Statistics and Computing 17 (4): 395–416.

Ma, Hanhui, Li-Chun Tu, Ardalan Naseri, Maximiliaan Huisman, Shaojie Zhang, David Grunwald, and Thoru Pederson. 2016. “Multiplexed Labeling of Genomic Loci with dCas9 and Engineered sgRNAs Using CRISPRainbow.” Nature Biotechnology 34 (5): 528–30.

Martins, André L., Ninad M. Walavalkar, Warren D. Anderson, Chongzhi Zang, and Michael J. Guertin. 2018. “Universal Correction of Enzymatic Sequence Bias Reveals Molecular Signatures of protein/DNA Interactions.” Nucleic Acids Research 46 (2): e9.

Mavrothalassitis, G., and J. Ghysdael. 2000. “Proteins of the ETS Family with Transcriptional Repressor Activity.” Oncogene 19 (55): 6524–32.

Medzhitov, Ruslan, and Tiffany Horng. 2009. “Transcriptional Control of the Inflammatory Response.” Nature Reviews Immunology. https://doi.org/10.1038/nri2634.

Medzhitov, Ruslan, and Tiffany Horng. 2009. “Transcriptional Control of the Inflammatory Response.” Nature Reviews. Immunology 9 (10): 692–703.

Mellor, Andrew L., and David H. Munn. 2004. “IDO Expression by Dendritic Cells: Tolerance and Tryptophan Catabolism.” Nature Reviews. Immunology 4 (10): 762–74.

Merli, C., D. E. Bergstrom, J. A. Cygan, and R. K. Blackman. 1996. “Promoter Specificity Mediates the Independent Regulation of Neighboring Genes.” Genes & Development 10 (10): 1260–70.

Mestas, Javier, and Christopher C. W. Hughes. 2004. “Of Mice and Not Men: Differences between Mouse and Human Immunology.” Journal of Immunology 172 (5): 2731–38.

Mezan, Shaul, Reut Ashwal-Fluss, Rom Shenhav, Manuel Garber, and Sebastian Kadener. 2013. “Genome-Wide Assessment of Post-Transcriptional Control in the Fly Brain.” Frontiers in Molecular Neuroscience 6 (December): 49.

Mikkelsen, Tarjei S., Zhao Xu, Xiaolan Zhang, Li Wang, Jeffrey M. Gimble, Eric S. Lander, and Evan D. Rosen. 2010. “Comparative Epigenomic Analysis of Murine and Human Adipogenesis.” Cell 143 (1): 156–69.

Minnich, Martina, Hiromi Tagoh, Peter Bönelt, Elin Axelsson, Maria Fischer, Beatriz Cebolla, Alexander Tarakhovsky, Stephen L. Nutt, Markus Jaritz, and Meinrad Busslinger. 2016. “Multifunctional Role of the Transcription Factor Blimp-1 in Coordinating Plasma Cell Differentiation.” Nature Immunology 17 (3): 331–43.

Mokrani, H., O. Sharaf El dein, Z. Mansuroglu, and E. Bonnefoy. 2006. “Binding of YY1 to the Proximal Region of the Murine Beta Interferon Promoter Is Essential To Allow CBP Recruitment and K8H4/K14H3 Acetylation on the Promoter Region after Virus Infection.” Molecular and Cellular Biology. https://doi.org/10.1128/mcb.00420-06.

Moore, Jill E., Henry E. Pratt, Michael J. Purcaro, and Zhiping Weng. 2020. “A Curated Benchmark of Enhancer-Gene Interactions for Evaluating Enhancer-Target Gene Prediction Methods.” Genome Biology 21 (1): 17.

Moorthy, Sakthi D., Scott Davidson, Virlana M. Shchuka, Gurdeep Singh, Nakisa Malek-Gilani, Lida Langroudi, Alexandre Martchenko, Vincent So, Neil N. Macpherson, and Jennifer A. Mitchell. 2017. “Enhancers and Super-Enhancers

Have an Equivalent Regulatory Role in Embryonic Stem Cells through Regulation of Single or Multiple Genes." Genome Research 27 (2): 246–58.

Mouse Genome Sequencing Consortium, Robert H. Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F. Abril, Pankaj Agarwal, et al. 2002. "Initial Sequencing and Comparative Analysis of the Mouse Genome." Nature 420 (6915): 520–62.

Mumbach, Maxwell R., Adam J. Rubin, Ryan A. Flynn, Chao Dai, Paul A. Khavari, William J. Greenleaf, and Howard Y. Chang. 2016. "HiChIP: Efficient and Sensitive Analysis of Protein-Directed Genome Architecture." Nature Methods 13 (11): 919–22.

Mumbach, Maxwell R., Ansuman T. Satpathy, Evan A. Boyle, Chao Dai, Benjamin G. Gowen, Seung Woo Cho, Michelle L. Nguyen, et al. 2017. "Enhancer Connectome in Primary Human Cells Identifies Target Genes of Disease-Associated DNA Elements." Nature Genetics 49 (11): 1602–12.

NCBI Resource Coordinators. 2016. "Database Resources of the National Center for Biotechnology Information." Nucleic Acids Research 44 (D1): D7–19.

Necsulea, Anamaria, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C. Baker, Frank Grützner, and Henrik Kaessmann. 2014. "The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods." Nature, January. https://doi.org/10.1038/nature12943.

Netea, Mihai G. 2013. “Training Innate Immunity: The Changing Concept of Immunological Memory in Innate Host Defence.” European Journal of Clinical Investigation 43 (8): 881–84.

Nishizaki, Sierra S., and Alan P. Boyle. 2017. “Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms.” Trends in Genetics: TIG 33 (1): 34–45.

Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” Nature 485 (7398): 381–85.

Nott, Alexi, Inge R. Holtman, Nicole G. Coufal, Johannes C. M. Schlachetzki, Miao Yu, Rong Hu, Claudia Z. Han, et al. 2019. “Brain Cell Type-Specific Enhancer-Promoter Interactome Maps and Disease-Risk Association.” Science 366 (6469): 1134–39.

O’Donnell, Amanda, Shen-Hsi Yang, and Andrew D. Sharrocks. 2008. “MAP Kinase-Mediated c-Fos Regulation Relies on a Histone Acetylation Relay Switch.” Molecular Cell 29 (6): 780–85.

Odom, Duncan T., Robin D. Dowell, Elizabeth S. Jacobsen, William Gordon, Timothy W. Danford, Kenzie D. MacIsaac, P. Alexander Rolfe, Caitlin M. Conboy, David K. Gifford, and Ernest Fraenkel. 2007. “Tissue-Specific Transcriptional Regulation Has Diverged Significantly between Human and Mouse.” Nature Genetics 39 (6): 730–32.

Ong, Chin-Tong, and Victor G. Corces. 2011. “Enhancer Function: New Insights into the Regulation of Tissue-Specific Gene Expression.” Nature Reviews. Genetics 12 (4): 283–93.

Ovsiy, Ilja, Vladimir Riabov, Ioannis Manousaridis, Julia Michel, Kondaiah Moganti, Shuiping Yin, Tengfei Liu, et al. 2017. “IL-4 Driven Transcription Factor FoxQ1 Is Expressed by Monocytes in Atopic Dermatitis and Stimulates Monocyte Migration.” Scientific Reports 7 (1): 16847.

Palikyras, Spiros, and Argyris Papantonis. 2019. “Modes of Phase Separation Affecting Chromatin Regulation.” Open Biology 9 (10): 190167.

Panne, Daniel, Tom Maniatis, and Stephen C. Harrison. 2007. “An Atomic Model of the Interferon-β Enhanceosome.” Cell 129 (6): 1111–23.

Parnas, Oren, Marko Jovanovic, Thomas M. Eisenhaure, Rebecca H. Herbst, Atray Dixit, Chun Jimmie Ye, Dariusz Przybylski, et al. 2015. “A Genome-Wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks.” Cell 162 (3): 675–86.

Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. “In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences.” Nature 444 (7118): 499–502.

Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. “Enhancers: Five Essential Questions.” Nature Reviews. Genetics 14 (4): 288–95.

Perry, Michael W., Alistair N. Boettiger, Jacques P. Bothma, and Michael Levine. 2010. “Shadow Enhancers Foster Robustness of Drosophila Gastrulation.” Current Biology: CB 20 (17): 1562–67.

Phan, Dillon, Chien-Jui Cheng, Matthew Galfione, Funda Vakar-Lopez, James Tunstead, Nancy E. Thompson, Richard R. Burgess, Sonia M. Najjar, Li-Yuan Yu-Lee, and Sue-Hwa Lin. 2004. “Identification of Sp2 as a Transcriptional Repressor of Carcinoembryonic Antigen-Related Cell Adhesion Molecule 1 in Tumorigenesis.” Cancer Research 64 (9): 3072–78.

Phanstiel, Douglas H., Kevin Van Bortle, Damek Spacek, Gaelen T. Hess, Muhammad Saad Shamim, Ido Machol, Michael I. Love, Erez Lieberman Aiden, Michael C. Bassik, and Michael P. Snyder. 2017. “Static and Dynamic DNA Loops Form AP-1-Bound Activation Hubs during Macrophage Development.” Molecular Cell 67 (6): 1037–48.e6.

Phillips, Theresa, and Others. 2008. “The Role of Methylation in Gene Expression.” Nature Education 1 (1): 116.

Pique-Regi, Roger, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. 2011. “Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data.” Genome Research 21 (3): 447–55.

Ponjavic, J., C. P. Ponting, and G. Lunter. 2007. “Functionality or Transcriptional Noise? Evidence for Selection within Long Noncoding RNAs.” Genome Research 17 (5): 556–65.

Pruitt, Kim D., Tatiana Tatusova, Garth R. Brown, and Donna R. Maglott. 2012. “NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy.” Nucleic Acids Research 40 (Database issue): D130–35.

Qiao, Yichun, Chiou-Nan Shiue, Jian Zhu, Ting Zhuang, Philip Jonsson, Anthony P. H. Wright, Chunyan Zhao, and Karin Dahlman-Wright. 2015. “AP-1-Mediated Chromatin Looping Regulates ZEB2 Transcription: New Insights into TNFα-Induced Epithelial-Mesenchymal Transition in Triple-Negative Breast Cancer.” Oncotarget 6 (10): 7804–14.

Quach, Bryan, and Terrence S. Furey. 2017. “DeFCoM: Analysis and Modeling of Transcription Factor Binding Sites Using a Motif-Centric Genomic Footprinter.” Bioinformatics 33 (7): 956–63.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” Bioinformatics 26 (6): 841–42.

Quinodoz, Sofia A., Noah Ollikainen, Barbara Tabak, Ali Palla, Jan Marten Schmidt, Elizabeth Detmar, Mason M. Lai, et al. 2018. “Higher-Order Inter-Chromosomal Hubs Shape 3D Genome Organization in the Nucleus.” Cell 174 (3): 744–57.e24.

Rabani, Michal, Joshua Z. Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, et al. 2011. “Metabolic Labeling of RNA Uncovers Principles of RNA Production and Degradation Dynamics in Mammalian Cells.” Nature Biotechnology 29 (5): 436–42.

Rabani, Michal, Raktima Raychowdhury, Marko Jovanovic, Michael Rooney, Deborah J. Stumpo, Andrea Pauli, Nir Hacohen, et al. 2014. “High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies.” Cell 159 (7): 1698–1710.

Rada-Iglesias, Alvaro, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. 2011. “A Unique Chromatin Signature Uncovers Early Developmental Enhancers in Humans.” Nature 470 (7333): 279–83.

Raghupathy, Narayanan, Kwangbom Choi, Matthew J. Vincent, Glen L. Beane, Keith S. Sheppard, Steven C. Munger, Ron Korstanje, Fernando Pardo-Manual de Villena, and Gary A. Churchill. 2018. “Hierarchical Analysis of RNA-Seq Reads Improves the Accuracy of Allele-Specific Expression.” Bioinformatics 34 (13): 2177–84.

Raj, Anil, Heejung Shim, Yoav Gilad, Jonathan K. Pritchard, and Matthew Stephens. 2015. “msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding.” PloS One 10 (9): e0138030.

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. “deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis.” Nucleic Acids Research 44 (W1): W160–65.

Ramsey, Matthew R., Lei He, Nicole Forster, Benjamin Ory, and Leif W. Ellisen. 2011. “Physical Association of HDAC1 and HDAC2 with p63 Mediates Transcriptional Repression and Tumor Maintenance in Squamous Cell Carcinoma.” Cancer Research 71 (13): 4373–79.

Reinhard, Christian, Dario Bottinelli, Baek Kim, and Jeremy Luban. 2014. “Vpx Rescue of HIV-1 from the Antiviral State in Mature Dendritic Cells Is Independent of the Intracellular Deoxynucleotide Concentration.” Retrovirology 11 (February): 12.

Rieder, Dietmar, Zlatko Trajanoski, and James G. McNally. 2012. “Transcription Factories.” Frontiers in Genetics 3 (October): 221.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” Nature 518 (7539): 317–30.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” Nature Biotechnology 29 (1): 24–26.

Rubin, Adam J., Brook C. Barajas, Mayra Furlan-Magaril, Vanessa Lopez-Pajares, Maxwell R. Mumbach, Imani Howard, Daniel S. Kim, et al. 2017. “Lineage-Specific Dynamic and Pre-Established Enhancer-Promoter Contacts Cooperate in Terminal Differentiation.” Nature Genetics 49 (10): 1522–28.

Sabari, Benjamin R., Alessandra Dall'Agnese, Ann Boija, Isaac A. Klein, Eliot L. Coffey, Krishna Shrinivas, Brian J. Abraham, et al. 2018. "Coactivator Condensation at Super-Enhancers Links Phase Separation and Gene Control." Science 361 (6400). https://doi.org/10.1126/science.aar3958.

Schmidt, Dominic, Michael D. Wilson, Benoit Ballester, Petra C. Schwalie, Gordon D. Brown, Aileen Marshall, Claudia Kutter, et al. 2010. "Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding." Science 328 (5981): 1036–40.

Schoenfelder, Stefan, and Peter Fraser. 2019. "Long-Range Enhancer–promoter Contacts in Gene Expression Control." Nature Reviews. Genetics 20 (8): 437–55.

Schoenfelder, Stefan, Biola-Maria Javierre, Mayra Furlan-Magaril, Steven W. Wingett, and Peter Fraser. 2018. "Promoter Capture Hi-C: High-Resolution, Genome-Wide Profiling of Promoter Interactions." Journal of Visualized Experiments: JoVE, no. 136 (June). https://doi.org/10.3791/57320.

Schuijs, Martijn J., Monique A. Willart, Karl Vergote, Delphine Gras, Kim Deswarte, Markus J. Ege, Filipe Branco Madeira, et al. 2015. "Farm Dust and Endotoxin Protect against Allergy through A20 Induction in Lung Epithelial Cells." Science 349 (6252): 1106–10.

Schulz, Laura C. 2010. "The Dutch Hunger Winter and the Developmental Origins of Health and Disease." Proceedings of the National Academy of Sciences of the United States of America.

Schwarzer, Wibke, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A. Fonseca, et al. 2017. “Two Independent Modes of Chromatin Organization Revealed by Cohesin Removal.” Nature 551 (7678): 51–56.

Setty, Manu, and Christina S. Leslie. 2015. “SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps.” PLoS Computational Biology 11 (5): e1004271.

Sheikh, Knvul. 2018. “Human Genome Project Celebrates 15th Anniversary.” Genome Your Health Is Personal. April 5, 2018. http://www.genomemag.com/2018/04/human-genome-project-celebrates-15th-anniversary/.

Shen, Li, Ningyi Shao, Xiaochuan Liu, and Eric Nestler. 2014. “Ngs.plot: Quick Mining and Visualization of next-Generation Sequencing Data by Integrating Genomic Databases.” BMC Genomics 15 (April): 284.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. “Transcriptional Enhancers: From Properties to Genome-Wide Predictions.” Nature Reviews. Genetics 15 (4): 272–86.

Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. 2006. “Nuclear Organization of Active and Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–on-Chip (4C).” Nature Genetics 38 (11): 1348–54.

Smale, Stephen T., Alexander Tarakhovsky, and Gioacchino Natoli. 2014. “Chromatin Contributions to the Regulation of Innate Immunity.” Annual Review of Immunology 32 (February): 489–511.

Smit, A. F., and A. D. Riggs. 1995. “MIRs Are Classic, tRNA-Derived SINEs That Amplified before the Mammalian Radiation.” Nucleic Acids Research 23 (1): 98–102.

Smit, Afa, R. Hubley, and P. Green. 2004. “RepeatMasker Open-3.0. 2004.” Seattle (WA): Institute for Systems Biology.

Snetkova, Valentina, and Jane A. Skok. 2018. “Enhancer Talk.” Epigenomics 10 (4): 483–98.

Song, Michael, Xiaoyu Yang, Xingjie Ren, Lenka Maliskova, Bingkun Li, Ian R. Jones, Chao Wang, et al. 2019. “Mapping Cis-Regulatory Chromatin Contacts in Neural Cells Links Neuropsychiatric Disorder Risk Variants to Target Genes.” Nature Genetics 51 (8): 1252–62.

Splinter, E. 2006. “CTCF Mediates Long-Range Chromatin Looping and Local Histone Modification in the Beta-Globin Locus.” Genes & Development. https://doi.org/10.1101/gad.399506.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” Proceedings of the National Academy of Sciences of the United States of America 102 (43): 15545–50.

Sung, Myong-Hee, Michael J. Guertin, Songjoon Baek, and Gordon L. Hager. 2014. “DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence.” Molecular Cell 56 (2): 275–85.

Thanos, D., and T. Maniatis. 1995. “Virus Induction of Human IFN Beta Gene Expression Requires the Assembly of an Enhanceosome.” Cell 83 (7): 1091–1100.

Thanos, Dimitris, and Tom Maniatis. 1995. “Virus Induction of Human IFNβ Gene Expression Requires the Assembly of an Enhanceosome.” Cell 83 (7): 1091–1100.

Tie, Feng, Rakhee Banerjee, Carl A. Stratton, Jayashree Prasad-Sinha, Vincent Stepanik, Andrei Zlobin, Manuel O. Diaz, Peter C. Scacheri, and Peter J. Harte. 2009. “CBP-Mediated Acetylation of Histone H3 Lysine 27 Antagonizes Drosophila Polycomb Silencing.” Development 136 (18): 3131–41.

Toshchakov, Vladimir, Bryan W. Jones, Pin-Yu Perera, Karen Thomas, M. Joshua Cody, Shuling Zhang, Bryan R. G. Williams, et al. 2002. “TLR4, but Not TLR2, Mediates IFN-Beta-Induced STAT1alpha/beta-Dependent Gene Expression in Macrophages.” Nature Immunology 3 (4): 392–98.

Tripodi, Ignacio J., Mary A. Allen, and Robin D. Dowell. 2018. “Detecting Differential Transcription Factor Activity from ATAC-Seq Data.” Molecules 23 (5). https://doi.org/10.3390/molecules23051136.

Ulitsky, Igor. 2016. “Evolution to the Rescue: Using Comparative Genomics to Understand Long Non-Coding RNAs.” Nature Reviews. Genetics 17 (10): 601–14.

Vandenbon, Alexis, Yutaro Kumagai, Mengjie Lin, Yutaka Suzuki, and Kenta Nakai. 2018. “Waves of Chromatin Modifications in Mouse Dendritic Cells in Response to LPS Stimulation.” Genome Biology 19 (1): 138.

Vierbuchen, Thomas, Emi Ling, Christopher J. Cowley, Cameron H. Couch, Xiaofeng Wang, David A. Harmin, Charles W. M. Roberts, and Michael E. Greenberg. 2017. “AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection.” Molecular Cell 68 (6): 1067–82.e12.

Vierstra, Jeff, Eric Rynes, Richard Sandstrom, Miaohua Zhang, Theresa Canfield, R. Scott Hansen, Sandra Stehling-Sun, et al. 2014. “Mouse Regulatory DNA Landscapes Reveal Global Principles of Cis-Regulatory Evolution.” Science , November, 1246426.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. “Enhancer Evolution across 20 Mammalian Species.” Cell 160 (3): 554–66.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. “ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers.” Nature 457 (7231): 854–58.

Wang, Hongjie, Guiting Lin, and Zhiwen Zhang. 2007. “ATF5 Promotes Cell Survival through Transcriptional Activation of Hsp27 in H9c2 Cells.” Cell Biology International 31 (11): 1309–15.

Wang, Jie, Jiali Zhuang, Sowmya Iyer, Xinying Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, et al. 2012. “Sequence Features and Chromatin Structure around the Genomic Regions Bound by 119 Human Transcription Factors.” Genome Research 22 (9): 1798–1812.

Wang, Ting, Jue Zeng, Craig B. Lowe, Robert G. Sellers, Sofie R. Salama, Min Yang, Shawn M. Burgess, Rainer K. Brachmann, and David Haussler. 2007. “Species-Specific Endogenous Retroviruses Shape the Transcriptional Network of the Human Tumor Suppressor Protein p53.” Proceedings of the National Academy of Sciences 104 (47): 18613–18.

Washietl, Stefan, Manolis Kellis, and Manuel Garber. 2014. “Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals.” Genome Research, March. https://doi.org/10.1101/gr.165035.113.

Watatani, Yujiro, Natsumi Kimura, Yusuke I. Shimizu, Itsuka Akiyama, Daijuro Tonaki, Hidenori Hirose, Shigeru Takahashi, and Yuji Takahashi. 2007. “Amino Acid Limitation Induces Expression of ATF5 mRNA at the Post-Transcriptional Level.” Life Sciences 80 (9): 879–85.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014.

“Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.” Cell 158 (6): 1431–43.

Wen, Andy Y., Kathleen M. Sakamoto, and Lloyd S. Miller. 2010. “The Role of the Transcription Factor CREB in Immune Function.” Journal of Immunology 185 (11): 6413–19.

Whyte, Warren A., David A. Orlando, Denes Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong Ihn Lee, and Richard A. Young. 2013. “Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.” Cell 153 (2): 307–19.

Wingender, Edgar, Torsten Schoeps, and Jürgen Dönitz. 2013. “TFClass: An Expandable Hierarchical Classification of Human Transcription Factors.” Nucleic Acids Research 41 (Database issue): D165–70.

Xie, Shiqi, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. 2017. “Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells.” Molecular Cell 66 (2): 285–99.e5.

Xu, Tianlei, Xiaoqi Zheng, Ben Li, Peng Jin, Zhaohui Qin, and Hao Wu. 2018. “A Comprehensive Review of Computational Prediction of Genome-Wide Features.” Briefings in Bioinformatics, November. https://doi.org/10.1093/bib/bby110.

Yamaoka, K., T. Otsuka, H. Niiro, Y. Arinobu, Y. Niho, N. Hamasaki, and K. Izuhara. 1998. “Activation of STAT5 by Lipopolysaccharide through Granulocyte-

Macrophage Colony-Stimulating Factor Production in Human Monocytes." Journal of Immunology 160 (2): 838–45.

Yin, Lei, and Mitchell A. Lazar. 2005. "The Orphan Nuclear Receptor Rev-Erbalpha Recruits the N-CoR/histone Deacetylase 3 Corepressor to Regulate the Circadian Bmal1 Gene." Molecular Endocrinology 19 (6): 1452–59.

Yokoshi, Moe, Kazuma Segawa, and Takashi Fukaya. 2020. "Visualizing the Role of Boundary Elements in Enhancer-Promoter Communication." Molecular Cell, February. https://doi.org/10.1016/j.molcel.2020.02.007.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters." Omics: A Journal of Integrative Biology 16 (5): 284–87.

Yuan, L. W., and J. E. Gambee. 2001. "Histone Acetylation by p300 Is Involved in CREB-Mediated Transcription on Chromatin." Biochimica et Biophysica Acta 1541 (3): 161–69.

Yue, Feng, Yong Cheng, Alessandra Breschi, Jeff Vierstra, Weisheng Wu, Tyrone Ryba, Richard Sandstrom, et al. 2014. "A Comparative Encyclopedia of DNA Elements in the Mouse Genome." Nature 515 (7527): 355–64.

Yukselen, Onur, Osman Turkyilmaz, Ahmet Rasit Ozturk, Manuel Garber, and Alper Kucukural. 2019. "DolphinNext: A Graphical User Interface for Creating, Deploying and Executing Nextflow Pipelines." bioRxiv. https://doi.org/10.1101/689539.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” Genome Biology 9 (9): R137.

Zhao, Zhihu, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, et al. 2006. “Circular Chromosome Conformation Capture (4C) Uncovers Extensive Networks of Epigenetically Regulated Intra- and Interchromosomal Interactions.” Nature Genetics 38 (11): 1341–47.

Zheng, Meizhen, Simon Zhongyuan Tian, Daniel Capurso, Minji Kim, Rahul Maurya, Byoungkoo Lee, Emaly Piecuch, et al. 2019. “Multiplex Chromatin Interactions with Single-Molecule Precision.” Nature 566 (7745): 558–62.

Zhou, Vicky W., Alon Goren, and Bradley E. Bernstein. 2011. “Charting Histone Modifications and the Functional Organization of Mammalian Genomes.” Nature Reviews. Genetics 12 (1): 7–18.

Zhu, Yizhou, Cagdas Tazearslan, and Yousin Suh. 2017. “Challenges and Progress in Interpretation of Non-Coding Genetic Variants Associated with Human Disease.” Experimental Biology and Medicine 242 (13): 1325–34.